This is a digital copy of a book that was preserved for generations on library shelves before it was carefully scanned by Google as part of a project to make the world's books discoverable online.

It has survived long enough for the copyright to expire and the book to enter the public domain. A public domain book is one that was never subject to copyright or whose legal copyright term has expired. Whether a book is in the public domain may vary country to country. Public domain books are our gateways to the past, representing a wealth of history, culture and knowledge that's often difficult to discover.

Marks, notations and other marginalia present in the original volume will appear in this file - a reminder of this book's long journey from the publisher to a library and finally to you.

Usage guidelines

Google is proud to partner with libraries to digitize public domain materials and make them widely accessible. Public domain books belong to the public and we are merely their custodians. Nevertheless, this work is expensive, so in order to keep providing this resource, we have taken steps to prevent abuse by commercial parties, including placing technical restrictions on automated querying.

We also ask that you:

+ *Make non-commercial use of the files* We designed Google Book Search for use by individuals, and we request that you use these files for personal, non-commercial purposes.

+ *Refrain from automated querying* Do not send automated queries of any sort to Google's system: If you are conducting research on machine translation, optical character recognition or other areas where access to a large amount of text is helpful, please contact us. We encourage the use of public domain materials for these purposes and may be able to help.

+ *Maintain attribution* The Google "watermark" you see on each file is essential for informing people about this project and helping them find additional materials through Google Book Search. Please do not remove it.

+ *Keep it legal* Whatever your use, remember that you are responsible for ensuring that what you are doing is legal. Do not assume that just because we believe a book is in the public domain for users in the United States, that the work is also in the public domain for users in other countries. Whether a book is still in copyright varies from country to country, and we can't offer guidance on whether any specific use of any specific book is allowed. Please do not assume that a book's appearance in Google Book Search means it can be used in any manner anywhere in the world. Copyright infringement liability can be quite severe.

About Google Book Search

Google's mission is to organize the world's information and to make it universally accessible and useful. Google Book Search helps readers discover the world's books while helping authors and publishers reach new audiences. You can search through the full text of this book on the web at `http://books.google.com/`

FOR PRIVATE CIRCULATION ONLY.

HA
29
.B79

B 1,473,460

The Measurement of Groups and Series:

A COURSE OF LECTURES

BY

A. L. BOWLEY, M.A., F.S.S.,

APPOINTED TEACHER OF STATISTICS AT THE SCHOOL OF ECONOMICS (UNIVERSITY OF LONDON).

DELIVERED AT THE

INSTITUTE OF ACTUARIES, STAPLE INN HALL,

During the Session 1902-1903.

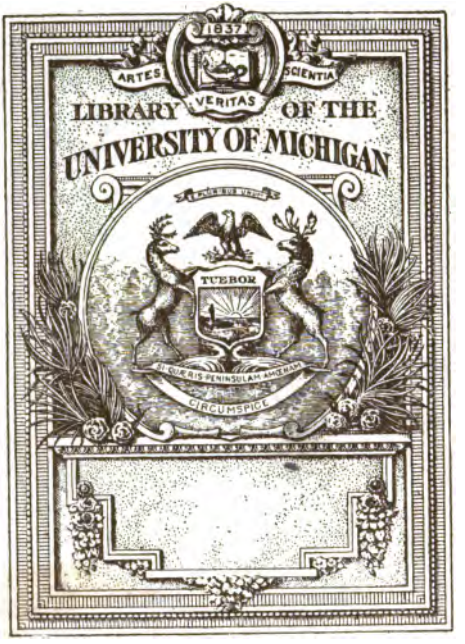
LONDON:

CHARLES AND EDWIN LAYTON,

56, FARRINGDON STREET, E.C.

1903.

Price 2/6.


1837
ARTES
SCIENTIA
VERITAS
LIBRARY OF THE
UNIVERSITY OF MICHIGAN
TUEBOR
CIRCUMSPICE

HA
29
.B79

The Measurement of Groups and Series:

A COURSE OF LECTURES

BY

ARTHUR L. BOWLEY, M.A., F.S.S.,

APPOINTED TEACHER OF STATISTICS AT THE SCHOOL OF ECONOMICS (UNIVERSITY OF LONDON).

DELIVERED AT THE

Institute of Actuaries, Staple Inn Hall,

During the Session 1902-1903.

LONDON:

CHARLES AND EDWIN LAYTON

56, FARRINGDON STREET, E.C.

1903.

PREFACE.

THE present course of Lectures on the Measurement of Groups and Series deals with some of the most modern methods of statistical research. Interesting as they were to those who had the advantage of hearing them delivered, they will doubtless, when studied at leisure in printed form, prove even more interesting and useful.

These Lectures are the fifth of a Series originated in 1897, designed for the assistance of Actuarial Students in connection with matters not included in the official Text Books. Three of the Series deal with legal matters, and one with the subject of Stock Exchange Securities. The present course carries the range of topics into the field of mathematics, and it is hoped that courses of lectures may be hereafter provided dealing with other subjects, practical and theoretical, relating to those branches of knowledge which it is the province of the Institute of Actuaries to promote and encourage.

W. H.

24th February, 1903.

330131

TABLE OF CONTENTS.

	PAGE
MEASUREMENT OF GROUPS	1
Graphic Method	2
Histogram and Ogive	7
Averages	11
Standard Deviation and Modulus	19
Average Deviation	24
Probable Error	24
Measurement of Skewness	26
The Curve of Error	31
The Method of Least Squares	45
Fitting Formulæ to Observations	48
Uses of the Curve of Error	53
Construction of a Group from Samples	56
Correlation between Two Groups	61
The Coefficient of Correlation	64
Justification of the Formula	72
MEASUREMENT OF SERIES	74
Classification	75
Periodic Curves	75
Symptomatic Series	77
Correlation between Series...	82
Criterion of Significance of the Correlation Coefficient ...	88
Conclusion	90

NOTE.—In the following lectures I have made free use of those statistical methods and formulæ, which (though in many cases of recent origin) may now perhaps be regarded as common property; but I hope that I have not inadvertently quoted without reference, or misquoted, investigations or theories, which may be regarded as personal to any of the small body of statisticians working on the subject treated. The lectures had to be prepared both for delivery and for the Press at short notice in the midst of a busy session. This fact must be my apology for any obscureness, unnecessary repetition, or clumsiness of arrangement or expression, which may be found.

A. L. BOWLEY.

MEASUREMENT OF GROUPS.

FIRST LECTURE.

GENTLEMEN, it was with considerable diffidence that I undertook to lecture to members of the Society of Actuaries on a subject with which they may be presumed to be so familiar. When I was asked if I could undertake these lectures I had some difficulty in choosing a suitable subject, and then it occurred to me that my audience were probably concerned with the practical aspects of a question which I was chiefly considering from the theoretical point of view, and that it would therefore be most suitable if I endeavoured to lay before them some theoretical considerations on subjects which did not come in their ordinary course, but which were allied to the subjects which naturally come before them, and which are allied to those subjects on which I have spent a certain amount of time and attention.

Groups.

The first subject which I have selected is the measurement of a group, the characteristics of a group, and its representation. By a group I understand a number of persons or things each of which possesses a measurable characteristic, the group being arranged according to the magnitude of the characteristic. For example, if I have returns of the wages of a large number of people, and I group them according to their wages, saying how many are earning 20*s.* to 25*s.*, and so forth, I shall have such a group; or if I choose a section of the population and group them according to ages, I should have another group of the kind I am thinking of. The

remarks I shall have to make about groups will, I hope, be fairly general, and apply to a very large range of groups; but for convenience of illustration I shall confine myself to only a small number.

The particular group I am taking for discussion this evening is taken from the current Census, the numbers of married women in the county of York, on Census day, 1901, grouped according to their ages. In selecting a group for discussion it must be large enough and small enough. It must be sufficiently large to conceal individual peculiarities, or peculiarities of small sections; it must be sufficiently small to be homogeneous. Both these limits are relative. A group that is large enough for one purpose is too large for another purpose; and a group that is homogeneous for one is heterogeneous for another. The death rate of a whole country may be sufficient for certain comparisons, but for other comparisons you must subdivide according to districts and age. The size that has been selected must be kept in view before any arguments are based on the grouping and its measurement.

There are two main divisions of groups: those that are derived from exact observations, and those which may be regarded as samples of a larger group the whole of which has not been measured. For purposes of reference I am calling them Group α, when the observations are supposed to be correct, as, for example, the number of persons who are in receipt of a certain income: and Group β, where the numbers are estimates; for example, an estimate of the number of persons who may be expected to be in receipt of certain incomes ten years hence, from an investigation of some group now, or at some previous time. As regards Group α, our chief work will be to select some method of abbreviating, of describing in brief, each group; in the case of Group β, our work will be chiefly to criticise the correctness of the statements, and to find methods which are properly applicable for its correction if it is not exact, to measure its precision, and then afterwards to select some suitable method of abbreviating it.

The Graphic Method.

The two chief methods of abbreviating or investigating the characteristics of a group are the graphic method and the method of averages. The method of averages should

perhaps be referred to first; but, since the use of diagrams in explaining the meaning of averages is very considerable, I have thought it better to take the method of diagrams first. I have drawn out, in four different ways, the group already named, the number of married women in the county of York.

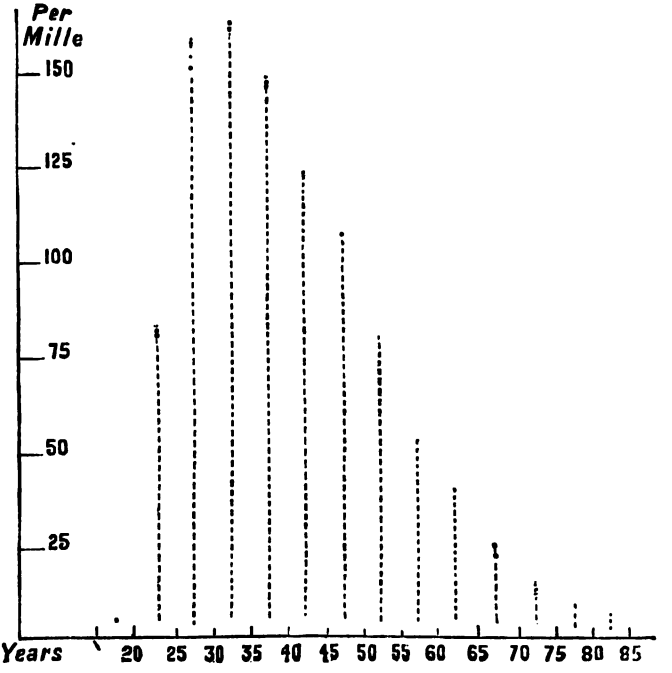
Ages of Wives present with their Husbands in the Registration County of York, 1901.

	No. per 1,000.			Per 1,000.
Between—			Not more than—	
15 and 16 years	·01		16 years old	·01
16 „ 17 „	·03		17 „	·04
17 „ 18 „	·2	5	18 „	·26
18 „ 19 „	1·2		19 „	1·5
19 „ 20 „	3·4		20 „	5
20 „ 21 „	8	83	21 „	13
21 „ 25 „	75		25 „	88
25 „ 30 „	157		30 „	245
30 „ 35 „	162		35 „	407
35 „ 40 „	147		40 „	554
40 „ 45 „	125		45 „	679
45 „ 50 „	105		50 „	784
50 „ 55 „	80		55 „	864
55 „ 60 „	55		60 „	919
60 „ 65 „	40		65 „	959
65 „ 70 „	22		70 „	981
70 „ 75 „	14		75 „	995
75 „ 80 „	4		80 „	999
Above 80 „	1			
	1,000			

Total number included, **610,505**.

There are shown in Diagrams I to IV, the numbers of married women in that county per thousand between these ages. The total of wives in the county of York living with their husbands was 610,000 odd. As is usual, the numbers are divided in years between the ages of 15 to 21, and after that in five-yearly groups. The first method of representing figures by diagram is to place a dot in a given vertical position for each person or item in question. This is indicated in Diagram I. The method is not very important and is perfectly obvious. I should only use it as a means of passing to another, if it were

I.—Point Diagram.



not that in those classes of measurement where the quantities are separated by a finite interval it is incorrect to use the methods shown in Diagrams II, III and IV. If one was entering the number of houses at particular rents in a town, where it might perhaps be supposed that the rent always jumped by as much as £2, one could represent properly the number of houses at each £2 mark, but there would be no house at the intermediate intervals, and it would be incorrect to proceed any further to such a curve as would lead one to suppose that the quantity dealt with was continuous. To take another example, the railway service from one town to another might be represented by a series of dots placed vertically over the time taken by the train, measured horizontally, but not by the following methods. If, however, the quantity is capable of continuous variation, such as age or height, or if by a slight extension of the meaning it may be regarded as being capable of continuous variation, such as income, we may proceed to the method of Diagram II.

In Diagram II rectangles are drawn whose heights are the same as the height of corresponding lines of dots in

II.—Area Diagram.

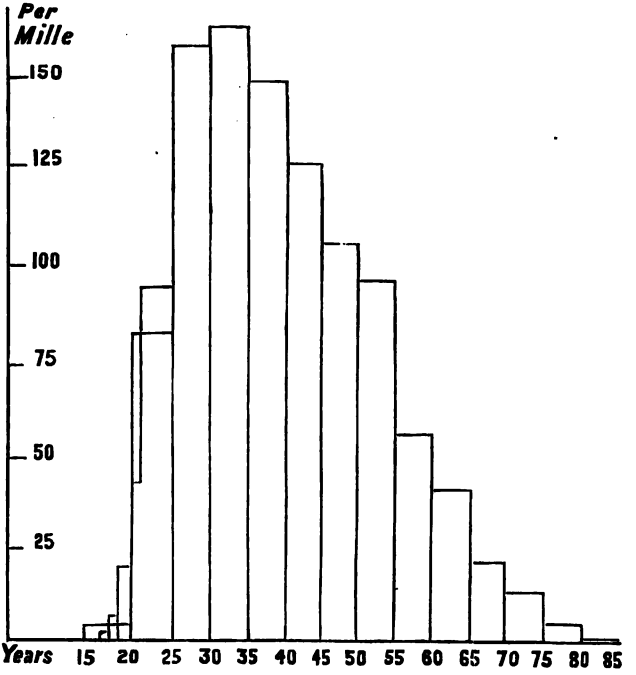


Diagram I, but the breadth is the unit of abscissæ, in this case five years. The areas can be regarded as representing a number of persons. The area of the whole space enclosed by the outer lines of the rectangles is on the scale chosen, 4 square inches, which represents the whole of the population considered; the breadth of each rectangle* is $\frac{1}{5}$ inch, and 1 inch squared represents 1 per cent.

Before we can go any further we have to make some assumption as to the distribution of persons within the five-yearly intervals selected. Even in my class, (*a*), when the observations are known to be correct, some assumption must be made as to distribution before proceeding further. If, for example, the correct set of measurements of the heights of a regiment was given, every soldier being measured correctly to the nearest $\frac{1}{10}$ of an inch, no correction would be required for actual mistakes, but before a continuous curve could be drawn passing from one $\frac{1}{10}$ inch to the next, some assumption must be made as to distribution of heights, *e.g.*, that progression was uniform between the given points. In the case

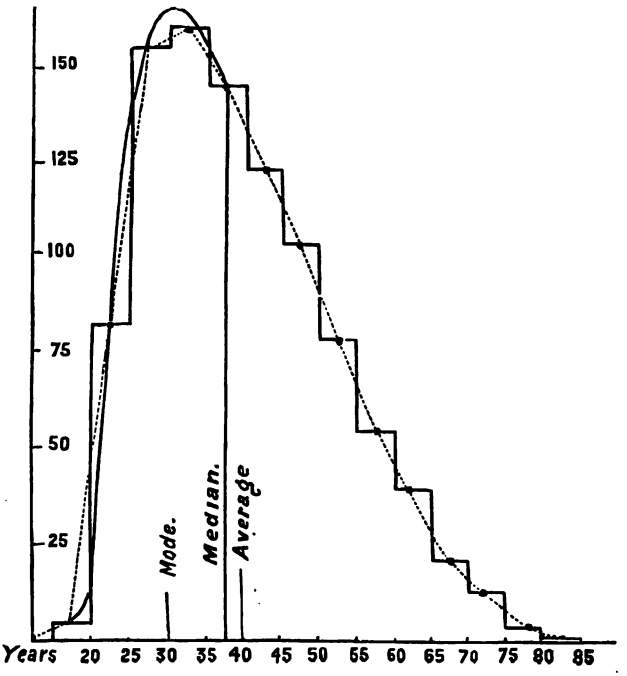
* The areas for ages below 25 are shown in more detail.

of observations (β) which are only samples, it is still more necessary to make some assumption as to continuity. In order to consider what assumption is proper, in the case in question, remember that the facts given exactly are the numbers of persons whose ages lie between certain limits; that is, we are given the area of the rectangle, or of the figure which replaces the rectangle on each unit of axis. What we have to suppose is that the ages are subdivided not merely into years, but into infinitesimal units of time; and we have to make some assumption for guiding us in passing from one of the given positions to the next. There are certain positions which give definitely the number of persons below 20 years, below 25 years, and so forth. We have to find the number of persons below 23, or any other assigned age. That is a quite familiar idea; but there are one or two things in connection with it which it is necessary to point out.

The Histogram and the Ogive.

If straight lines are drawn from the middle of each horizontal line in Diagram II to the middle of the next we get the dotted line in Diagram III (called a histogram).

III.—Histogram.



That is certain to be incorrect on two grounds. In the first place, the area bounded by the lines nearest the highest point is necessarily too small, for part of the area between the ages 30 and 35 is cut off by the dotted line, and nothing is placed instead of it. Before pointing out the other way in which the histogram is necessarily incorrect, we will pass on to Diagram IV, which is thus constructed. At the ages shown on the horizontal axis are drawn rectangles proportional to the number of persons below that age; then we get a continually ascending figure called an ogive, which is given as absolutely correct in group *a*, the points at the corners of the steps obtained in the figure being given by assumption. The problem comes to be to draw some line or curve from these fixed points that shall satisfy the conditions which we must assign. Now, it would be necessarily wrong to join these successive points by straight lines. If we take three corners, A, B, C, not in a straight line, we get a sharp angle at B. Introducing sharp angles there necessarily involves an error, for they indicate discontinuity at certain arbitrary points, which can correspond to no facts in nature. The angles obtained in the histogram are erroneous for similar reasons. If, as in group *a*, we are to suppose the observations to be correct, a continuous line must be drawn through all the given points which has no sharp angles in it, no sharp change of curvature. If, as in group β, we are not bound to assume that the observations are correct, the line may be drawn not passing through the points, but near them. Many groups may be represented with sufficient accuracy in rough work by drawing a freehand curve passing through the given points; it will be found that there is very little margin for drawing such a curve, if the rule is made that the curvature is never to be greater than necessary, that the direction is not changed more rapidly than is necessary to pass through the points. This condition, stated in mathematical language, supplies the main problem of interpolation.

In group β we are not bound to assume that the curve passes through all the points, and the question which is the best curve (drawn freehand or otherwise) *near* the points, needs the theory of probability for its discussion.

INTERPOLATION CURVE.

As regards group *a*, to which discussion may be confined. for the present, where the curve is to pass *through* all the

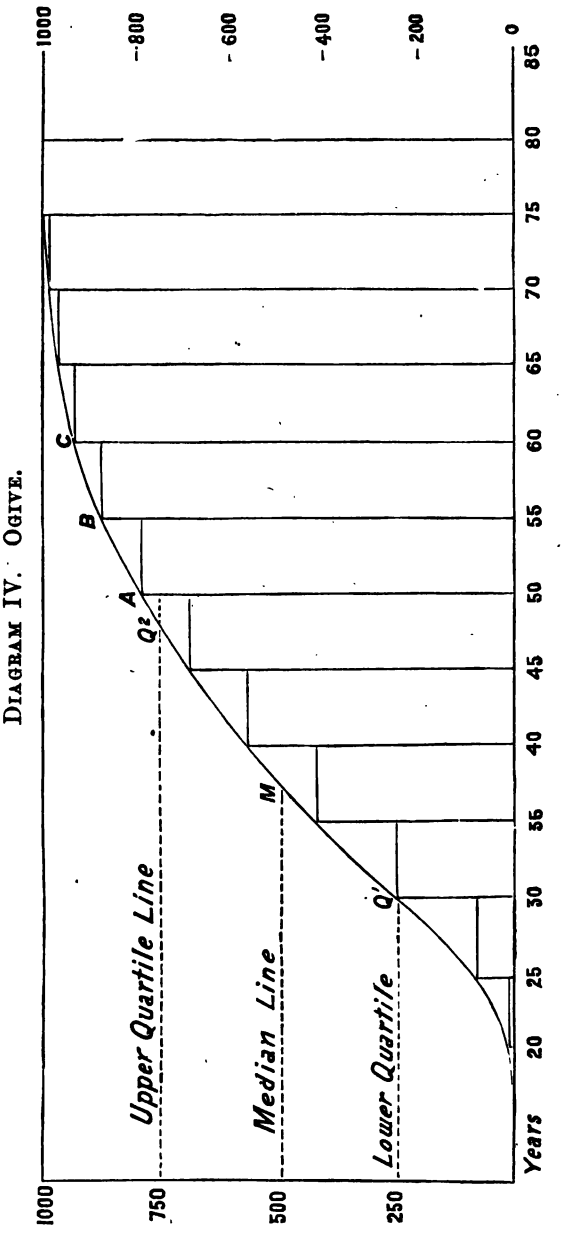


DIAGRAM IV. OGIVE.

points, I suggest the familiar method of interpolation by a parabolic formula. Take the equation $y = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots$ continued to as many terms as are convenient. In the group

under discussion it would be inexpedient to take more than 4 or 5 terms, because we are fitting a definite algebraic curve to irregular observations, and the law which underlies the observations may very well change if we take a larger period than 25 years. I have confined the work for this group to 5 terms, continuing that series to x^4.

Consider what conditions that curve satisfies. Stopping at the second term we have a straight line, which can only be made to pass through two points. So we have to start afresh at the second point, and thereby contradict the first assumption which we make, that the increase is not subject to violent changes, introducing an angle at any point. Introducing the second term x^2 we have a parabola, which can be made to pass through three points; the curve has continuous curvature, the third differential, and the third differences obtained from the values of y at three consecutive equidistant values of x, vanish; that is to say, there is no sudden change in curvature. The first differential measures the inclination of the line; the second measures the change of inclination, and if that is constant there is a constant change of inclination, but no sudden break. But we have no reason for assuming a constant change of inclination, and the curve which passes through three assigned points will not in general pass through the next point. We then proceed to include further terms. If we take the equation up to x^3 we can introduce a point of inflection, which we cannot do with the parabola. If we take the equation a step further we introduce two points of inflection, and it is unnecessary to go as far as the fifth term in a diagram like this. If we take 6 terms, the 6th differential and the 6th difference vanish, the 5th difference is constant, and there is no sudden break, and so on.

Now take the equation as far as the term x^4. We have 5 unknowns, and can determine them by assigning the condition that the curve shall pass exactly through 5 points on the diagram in question. I have calculated the equation of the curve which will pass exactly through the 5 points of the ogive, corresponding to 20, 25, 30, 35 and 40 years. The method of calculation is a good deal facilitated by the use of finite differences. Refer as origin for the abscissæ to age 20, take 5 years as unit, so that x is 1 at age 25, and write down the equations which naturally arise, taking the numbers from the last column of the table on p. 3.

$$5 = a_0$$
$$88 = a_0 + a_1 + a_2 + a_3 + a_4$$
$$245 = a_0 + a_1.2 + a_2.2^2 + a_3.2^3 + a_4.2^4$$
$$407 = a_0 + a_1.3 + a_2.3^2 + a_3.3^3 + a_4.3^4$$
$$554 = a_0 + a_1.4 + a_2.4^2 + a_3.4^3 + a_4.4^4.$$

It is easily shown that $a_4 = \Delta_0^4 \div 24$, $a_3 = \Delta_0^3 \div 6 - \Delta_0^4 \div 4$, $a_2 = \Delta_0^2 \div 2 - \Delta_0^3 \div 2 + \frac{11}{24}$ of Δ_0^4, and a_1 and a_0 are then easily calculated. In this case $\Delta_0^4 = 49$, $\Delta_0^3 = -69$, $\Delta_0^2 = 74$, $a_4 = 2\frac{1}{24}$, $a_3 = -23\frac{3}{4}$, $a_2 = 93\frac{3}{24}\frac{3}{4}$, $a_1 = 10\frac{3}{4}$.

Let $z = f(x)$ be the equation of the curve replacing the histogram, and $y = \mathrm{F}(x)$ the equation of the ogive. Then $y = \int_0^x z.dx$, and $\frac{dy}{dx} = z$. By means of these equations, the parabolic or smoothed curve now obtained from Diagram IV can be used to furnish values to replace the histogram of Diagram III. The same unit for x (five years) must, of course, be used in both cases. Thus, if $x = 1$ (25 years),

$$z = \frac{dy}{dx} = a_1 + 2a_2.1 + 3a_3.1^2 + 4a_4.1^3 = 135{\cdot}6.$$

At 30 years, $z = 166{\cdot}9$; at 35 years, $158{\cdot}8$.

The curve obtained in this way is shown in the continuous line in Diagram III; this curve satisfies the conditions that the areas standing on the 5-year bases, from 20 years to 45 years, should represent on the chosen scale the number of persons given by the original table, and that there should be no abrupt changes of curvature.

Since the curve has been chosen so as to satisfy the conditions for only five age periods, it will not necessarily satisfy any more; but in this case the curve merges into a straight line, which approximately fulfils the conditions till 65 years. If we need greater accuracy in later years, we should calculate new values for the a's and obtain a second curve, satisfying a new group of area conditions. If we needed to draw the whole curve accurately, we should have to devise a method of passing without a break of continuity from one such parabolic curve to the next; but, as it is, we only want means of obtaining specified points on the curve, and that can be done by choosing the special parabolic curve that is in the neighbourhood of the required points.

Averages.

The Mode.—At the highest point of the smooth curve in Diagram III, $\frac{dz}{dx}=0$; hence $\frac{d^2y}{dx^2}=0$ in the ogive for the same value of x. Thus, as is otherwise evident, the ogive is steepest and there is a point of inflexion, at that value of x which gives the greatest ordinate in Diagram III.

If $\frac{d^2y}{dx^2}=0$, we have $0=2a_2+6a_3x+12a_4x^2$, and

$$x=(-3a_3\pm\sqrt{9a_3^2-24a_2a_4})\div 12a_4,$$

where a_2, a_3, a_4 are given in terms of the differences above. Writing in these values, we have $x=2{\cdot}021$, which corresponds to 30·10 years.

If we had included a further term, a_5x^5, we should have a cubic to solve to determine x. If we had only gone as far as a_3x^3, we should have the equation $0=a_2+3a_3x$; that is, $x=1-\Delta_0^2\div\Delta_0^3$; but this formula is unsafe, unless the fourth differences of the original figures are approximately zero.

The equation taken as far as the x^4 term appears to me to be practically the best in the example we are discussing. If we choose the coefficients to satisfy the conditions starting from the age 25, we obtain 30·43 years as the position of the highest point. The discrepancy between this and the 30·10 years found from the parabola starting from the age 20 years, arises from the indeterminateness of the original figures. It seems best to take the value from the former curve, as the point then lies near the middle of the assigned values. We adopt then the age 30·10 years as the required age, and find that $y=166{\cdot}94$ at that age. The age so found is called the *mode* of the group; it is also called the position of *greatest density* and of the *maximum ordinate*.

The Median.—The abscissa of the point (M) where the ogive is cut by the horizontal line half way up the scale (from 0 to 1,000) is called the *median*. In the histogram, or the smoothed curve which replaces it, the vertical through the median divides the curve into equal areas. When the ogive is drawn, the median can at once be found graphically. To find it algebraically, take a parabolic equation as before, satisfied by five points lying near the median, obtain the coefficients as before, and put

$y=500$. One of the roots of the equation so obtained is the median. Starting at 30 years we have—

$$245+164\tfrac{1}{12}x+\tfrac{1}{8}x^2-3\tfrac{5}{12}x^3+\tfrac{3}{8}x^4=y=500,$$

and, solving by Horner's method, $x=1{\cdot}623$, so that the median age is $30+5x=38{\cdot}11$ (years).

The *quartiles* are the abscissæ of the points (Q_1, Q_2) where the horizontal lines one-quarter and three-quarters up the scale (from 0 to 1,000) cut the ogive. The vertical through these abscissæ in Diagram III would, together with the median vertical, divide the area into four equal parts. The quartiles can be found from one of the equations already written by putting $y=250$ and 750 successively, and solving for x; or they can be found graphically.

A rough method of finding these points, often sufficiently accurate, and saving a more laborious solution, is to assume that the parts of the ogive between the corners which contain the median are straight lines. There are 407 (per thousand) below 35 years; 93 out of the 35 to 40 year group, which contains 147, are to be taken to reach the median, which is on the hypothesis of a straight line ($35+\frac{93}{147}$ of 5) years, that is 38·17 years, a value differing little from that already obtained. The lower quartile by either method is 30·16 years, the upper 48·6 years.

We have now the following figures:—

Lower quartile ...	30·16	years.
Mode	30·10	„
Median	38·11	„
Arithmetic average ...	40·111	„
Upper quartile ...	48·6	„

The arithmetic average is calculated directly in the ordinary way, but is of little importance in such a group as this.

If a person is taken at random from this group, her most probable age is 30·1 years, the mode. It is as likely as not that she will be over 38·11 years, the median. It is as likely as not that she will be between 30·16 and 48·6 years. The chances are 3 to 1 against her being less than 30·16 years; 3 to 1 against her being over 48·6 years.

Other points can be obtained by dividing the group into ten equal parts, or one hundred equal parts. These are called the *deciles* and *percentiles* respectively.

The *mode* in such groups as this seems to be of special importance, as being the most probable value. It is entirely unaffected by the extremes. If the Census authorities had omitted all married women over 50 or all under 20 years in their enumeration the mode would be still in the same place. That is very important when we are dealing with inaccurate figures. In those curves which have a distinct mode, where the curve first tends upwards, reaches a height, and then comes down again without ever pausing or returning to a second height, and where there is a certain symmetry or similarity of distribution on either side of it, in such curves the mode is of special importance. If, on the other hand, you have a regular mountain range represented by your curve, the mode is probably of much less importance. If you have a single peak it is probably of importance. But though it is important in itself it is quite insufficient to describe the curve; it only tells you the position of one point; it does not tell you the steepness on either side, or the distance from there to any assigned point.

The *median* is affected by extremes to some extent. If the authorities had omitted all the married women over 50 the median would of course have been shifted, but not very much, for the area, which would have been left out at the extreme right, when halved and distributed in the neighbourhood of the median would be found to have caused only a very slight displacement of it. That can be verified from Diagram IV. To take an example which can be supplied by the diagram, suppose you omit all those beyond the 800 per-cent, which gives those above 55, then the line through the 400 would give the median, which a very rough measurement gives as 33 years. That is to say, the median has only been shifted five years by leaving out that immense number. If, instead of omitting these people over 55, the Census authorities had simply said, "Here is a married woman, obviously old, we do not know her age," and had entered her in that category, it would not have affected the median in the very least. The position of the extremes does not affect the median, only the number of instances. In the statistics with which I personally have to deal, often all that is known is this number. In this respect the median is very superior to the arithmetical average. The same applies to quartiles. If we do not know the exact *positions* of the

instances to the right or to the left of the quartiles it does not matter, provided we know the *numbers*. If you decide two quartiles and the median, you have three points on the ogive curve, three positions in the histogram, from which the whole can often be constructed with fair accuracy. The arithmetic average, or simply "the average," gives the abscissa of the centre of gravity of the group when plotted out as in Diagram III. The arithmetic average facilitates certain computations, but, in my experience, it is the least valuable of the means or averages which can be calculated; other people's experience may be different. It is very liable to error. If a part of the group is accidentally omitted the average is at once affected. If the numbers are correct and the positions not very far out, you would find by experiment that the arithmetic average has not moved much; but directly any numbers are left out, the arithmetic average is disturbed. But the reason I distrust the arithmetic average and do not advocate its use is, chiefly because it renders such fallacious arguments possible. If you are comparing one group with another, after a little interval the arithmetic average may have remained quite steady when the group has changed considerably, both the extremes having come in towards the mean; or it may shift when the group has not really changed its character, but only shifted its position a little. Any particular change of the arithmetic average may correspond to an infinite number of different kinds of change in the group; and it is very often pointed out that a certain group has changed, that something has improved because the arithmetic average has changed; whereas it is only shifting the relative positions of two groups which are not connected in reality.* If we have a perfectly homogeneous group, for instance, if with wage statistics, we deal with a set of men doing similar work and earning similar wages, a change in the arithmetic average is significant; but if we are dealing with a composite group composed of skilled and unskilled workmen, two homogeneous groups merged into one, the arithmetic average might increase either by the higher group ascending a little while the lower group went down nearly as far, or the other way about; or by a combination of those two things.

* These two sentences apply also to the median, but the present unfamiliarity of the term will suggest caution in using it; while, as a matter of fact, the arithmetic average is used very carelessly.

So the arithmetic average can never give definite information, and very often gives fallacious information. I have not time, and perhaps it is not necessary, to dwell upon this point, and refer to the correction factors for Urban death rates. The necessity of that method illustrates my meaning in saying that before an arithmetic average is used, it is necessary to make sure that the group is homogeneous.

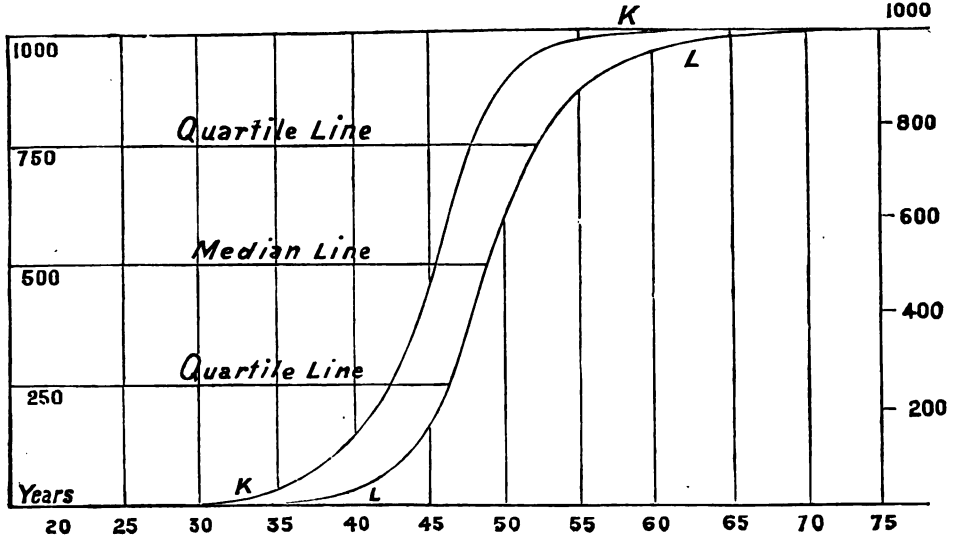
The quartiles and the median not only give the definite position of the median, but also a measurement, which serves to show how the curve is dispersed from its central position. The distance between the two quartiles, 18·4 years in this case, shows to some extent how the curve is dispersed from its central point. That I shall return to in giving other measurements of this dispersion.

If we were dealing with a group that did not give any such regular figure as this, a group to which the mode was certainly quite applicable, it would probably then be best not to attempt to draw any continuous curve at all, but to keep to such a diagram as that on page 5, and to calculate the deciles as accurately as possible. By making some simple assumptions as to continuity, it would be possible to calculate roughly the nine deciles, dividing the area into 10 equal parts, and enter them as a description of the group. I think that is the only method of satisfactorily representing an irregular group which cannot be divided into distinct homogeneous groups.

Comparison of Groups.

The ogive diagram lends itself more readily than any other to the comparison of the two groups. I have selected two groups, which one might wish to compare, from the same Census table, the husbands whose wives were between 45 and 50 years of age, and the wives whose husbands were between 40 and 45, which are represented by the lines LL and KK respectively; and I have calculated, by one method or the other, the mode, the median and the quartile of those groups. Thus, for instance, from the curve K, of all the wives whose husbands were between 45 and 50 years of age, as many were less than 45·5 years as were more than that; and similarly for the quartiles. The curves are very similar, the husband curve being four years to the right of the other. The method needs no further comment.

DIAGRAM V



KK Wives of husbands between 45 and 50 years.

LL Husbands of wives between 45 and 50 years.

K	Average	44·98	Median	45·5
L		49·48		48·5

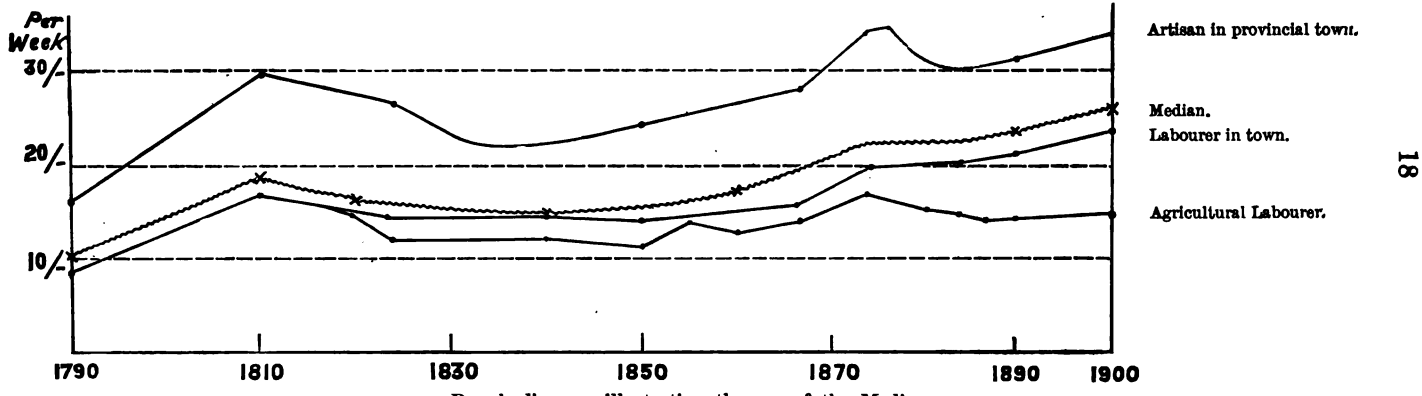
			Quartiles
K	Mode	46·35	42·54,
L		48·22	46·35, 52·5

Illustration of use of the Median.

I may take one example to illustrate the use of the median. The diagram on p. 18 represents the weekly wages, valuing everything that is paid in goods and not in money at an appropriate rate, of three classes of labourers in England, namely, Artisans in Provincial Towns, such as Birmingham, Agricultural Labourers—the average for the whole of England—and Labourers in the same towns from which the Artisans were selected. The figures are rather rough, and there is no material for making them exact; but I think the lines drawn represent with fair accuracy the course of wages; for if we once established the fact that all agricultural labourers are below the median, we have simply to count them and not enquire about their wages. And so if we establish the fact that any body of men is well above or well below the median, we have not to enquire into their wages, but simply to count them; and to find the median we have only to investigate more carefully the body of men whose wages are near the median; that is a comparatively easy task, because the body of men who are near to it are those whom we see any day in any ordinary industrial undertaking. The Census figures are bad for this purpose in 1902, and they were much worse in 1801; and there is a great deal of computation and guess-work in determining the position of the median at any time through the century. But it can be done within certain limits of accuracy where the task of determining the arithmetic average would be hopeless. When we have determined the median and trace out the positions for 110 years we have a much more interesting and exact piece of information than if we had made use of the arithmetic average. We have the wage of that man who is half way up the skilled wage earners; but if we give the arithmetic average it will carry us no further; it is simply a numerical quotient. The line in the diagram is drawn through the estimated positions of the median for all male adult wage earners in the United Kingdom, at selected dates. These figures are rough, and should not be quoted without verification. The only ones calculated are those with a dot or cross in the figure; intermediate lines are interpolated.

Diagram VI.

Wages in England in the XIXth Century.



Rough diagram illustrating the use of the Median.

MEASUREMENT OF GROUPS.

SECOND LECTURE.

The Standard Deviation and the Modulus.

THE methods I have employed so far for determining the median and the mode, together with the ordinary method of determining the arithmetic average, together also with the quartiles and deciles, give a series of definite quantities connected with the curve. Each of these quantities—the mode and the median—performs the function of an average; that is to say, that number by itself gives briefly one of the most important positions, one of the most important characteristics of the whole curve. But no one of these quantities gives sufficient information to enable us to reconstruct the curve or to describe it completely. It is true that if we have given the nine deciles, including the median, we have nine points on a continuous curve, and in general it is possible to construct it with reasonable accuracy. But if we only have the mode, or only the median, we have not enough to construct the curve. My object, then, is to develop one or more methods of calculating other quantities related to the group, which will enable us to complete or amend the description of the group, as given simply by one of the averages.

We will always suppose that the group is described in relation to a horizontal axis OX, and may be of any nature about the axis. What we have found so far in the median or the mode is one point on that group, one position on that axis—in the case of the mode the position under the highest point—in the case of the median the position, the line through which divides the curve into two equal areas—in the case of the average it is the abscissa of the centre of gravity. I have now to find a second quantity which will enable us to describe or determine the shape of the curve when you are given this one position on it. The method I am going to take is independent of any assumed shape of the curve, and it is

applicable to both the groups to which I referred on page 2, the group which is supposed to be an accurate representation of the facts, and that which represents only samples of a larger group whose observation is not completely made. I have first to describe the well-known method of calculating the deviations from the average, and then to pass on to find the average deviation, the average square of the deviation, and the average cube of deviation. Let there be n observations represented by $x_1, x_2 \ldots x_n$; let $\bar{x}$ be the abscissa of the centre of gravity; then $\bar{x}$ is the average of the group, the sum of the x's divided by n, their number. From each of the x's subtract the abscissa of the centre of gravity; thus $x_1-\bar{x}$, $x_2-\bar{x}, \ldots x_n-\bar{x}$. Those are the deviations of the observations from their average. In some connections they are called the errors from the average, but I shall adopt the word "deviation" in every case. In the first place it is to be noticed that the sum of the deviations is necessarily zero; for

$$\Sigma_1^n(x-\bar{x}) = \Sigma_1^n x - n\bar{x} = 0.$$

The sum of the squares of the deviations is $\Sigma_1^n(x-\bar{x})^2$, and $\frac{1}{n}\Sigma_1^n(x-\bar{x})^2$ is the mean square of the deviations, which is otherwise called the second moment of the deviations, about the origin in this case. The word moment is from a dynamical analogy; it is used in this connection by Professor Karl Pearson.

The following notation is adopted. The moments measured about the origin are written $\mu_1', \mu_2' \ldots$, about the centre of gravity $\mu_1, \mu_2 \ldots$, so that

$$\mu_1' = \frac{1}{n}\Sigma x,\ \mu_2' = \frac{1}{n}\Sigma x^2,\ \mu_3' = \frac{1}{n}\Sigma x^3 \ldots,$$

and

$$\mu_1 = \frac{1}{n}\Sigma(x-\bar{x}) = \frac{1}{n}\Sigma x - \bar{x} = 0,$$

$$\mu_2 = \frac{1}{n}\Sigma(x-\bar{x})^2 = \frac{1}{n}\Sigma x^2 - \frac{2\bar{x}}{n}\Sigma x + \bar{x}^2 = \mu_2' - \bar{x}^2$$

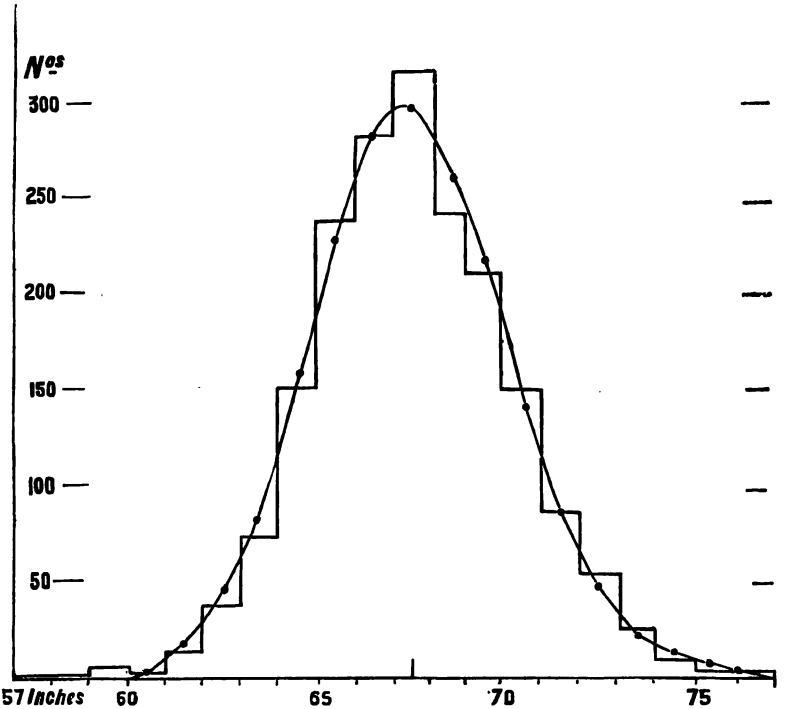
$$\mu_3 = \frac{1}{n}\Sigma(x-\bar{x})^3 = \frac{1}{n}\Sigma x^3 - \frac{3\bar{x}}{n}\Sigma x^2 + \frac{3\bar{x}^2}{n}\Sigma x - \bar{x}^3 = \mu_3' - 3\bar{x}\mu_2' + 2\bar{x}^3.$$

The quantities we need are not the moments about an arbitrary origin, but the moments about the centre of gravity. But it is far easier to calculate the moments about an arbitrary origin than to obtain those about the centre of gravity by the above formulæ.

Diagram VII.

Heights of 1,935 *Persons.*

Observations and Curve of Error.



$x+57\frac{1}{2}$	y	xy	xy^2	xy^3	$x+57\frac{1}{2}$	y	xy	xy^2	xy^3
Inches	Instances				Inches	Instances			
57—58	1	0	0	0	67—68	321	3,210	32,100	321,000
58—59	1	1	1	1	68—69	245	2,695	29,645	326,095
59—60	6	12	24	48	69—70	213	2,556	30,672	368,064
60—61	4	12	36	108	70—71	152	1,976	25,688	333,944
61—62	15	60	240	960	71—72	88	1,232	17,248	241,472
62—63	39	195	975	4,875	72—73	55	825	12,375	185,625
63—64	74	444	2,664	15,984	73—74	26	416	6,656	106,496
64—65	153	1,071	7,497	52,479	74—75	9	153	2,601	44,217
65—66	243	1,944	15,552	124,416	75—76	1	18	324	5,832
66—67	288	2,592	23,328	209,952	76—77	1	19	361	6,859
	824	6,331			Total ..	1,935	19,481	207,987	2,348,427

$n = 1{,}935.$

$\Sigma y = 19{,}431.$

$\bar{x} = \frac{\Sigma xy}{n} = 10{\cdot}0419.$ Average 67·54 inches.

$\Sigma y^2 = 207{,}987.$

$\frac{\Sigma y^2}{n} = 107{\cdot}487 = \mu'_2.$

$\frac{\Sigma y^3}{n} = 1{,}213{\cdot}66 = \mu'_3.$

ferred to average—

$\mu_2 = \mu'_2 - \bar{x}^2 = 6{\cdot}647$ (2nd moment).

$m_2 = \mu_2 - 1\frac{1}{12} = 6{\cdot}564$ (2nd moment corrected).

σ (standard deviation) $= \sqrt{m_2} = 2{\cdot}562$ ins.

c (modulus) $= \sqrt{2m_2} = 3{\cdot}623$ inches.

$\mu_3 = \mu'_3 - 3\mu'_2\bar{x} + 2\bar{x}^3 = 1{,}213{\cdot}66 - 3{,}238{\cdot}12 + 2{,}025{\cdot}24 = {\cdot}78$ (3rd moment).

$j = \frac{\mu_3}{c^3} = +{\cdot}016$ (skewness from moments).

$j = +{\cdot}06$ (skewness from observations).

η (mean deviation from average) $= 2{\cdot}01$ (ins.).

$\frac{c}{\sqrt{\pi}} = \frac{3{\cdot}63}{1{\cdot}772} = 2{\cdot}05$ (inches).

Median = average $- \frac{1}{2}jc = 67{\cdot}47$ (inches .

Mode = average $- jc = 67{\cdot}32$ (inches).

By parabolic interpolation—Mode, 67·303 inches. Median, 67·566 inches

It will now be convenient to follow the figures on the table and diagram adjoining. The figures are taken from the report of the Anthropometrical Committee of the British Association, in 1881. They are selected merely as being a convenient group by which to explain the calculation of these moments. The heights of 1,935 persons were given as between certain inches, between 57 and 58 inches, as under the column headed $x+57\frac{1}{2}$. I take the origin at $57\frac{1}{2}$ inches, and the abscissæ for the successive groups are 1, 2, 3 .... 20. The number of instances in these various groups are those given in the second column, under the letter y; one person under 58 inches, one between 58 and 59 inches, and so on. The instances in this case occur in groups, and we are not able to separate them by means of the data, hence each deviation will occur in most cases more than once. Thus, a deviation shown between 64 and 65 inches occurs 153 times. Instead of adding the x's simply to obtain the deviation we multiply each deviation by the number y, the number of times it occurs, and so obtain the third column xy, whose sum is 19,431, which is the first moment about the origin. The sum of the deviations is to be divided by 1,935, the total number of deviations, to give the first moment, namely, 10·042, and this gives the position of the centre of gravity measured from the origin, $57\frac{1}{2}$ inches. The columns under xy^2 and xy^3 require no explanation. The totals 207,987 and 2,348,427 are divided by n, giving 107·5 and 1,213, μ'_2 and μ'_3 in the notation adopted. It now remains to reduce these moments about the origin to the moments about the centre of gravity, by means of the formulæ given above.

The practical simplicity of evaluating the moments by this method arises from the fact that we are dealing in the x's with a series of numbers ascending in uniform order, 1 to 20, and that the whole arithmetic computation is very simple and very easily checked, whereas if we proceed on the direct method of writing down the position of the centre of gravity, which will naturally not be an exact number, each of the deviations will introduce as many decimal places as are kept in our calculation; and the squaring and cubing will be very arduous, and we have no ready means of checking our results. It is therefore worth while to take the formula and choose our origin so as to give the least arithmetic work and obtain the second and third moments indirectly. There is a

small correction to be made for the moments so calculated for the second, fourth, and other even moments. I will deal only with the second. It is to be observed that in the whole calculation it is assumed that all the persons in a particular group are exactly at the middle of that group, *e.g.*, that 153 persons in the 64 to 65 inches have the height exactly 64½ inches. It is obvious that that will not be the case, and it is easily seen that that will introduce a definite error in the calculation of the second moment. For if we take one of these groups in particular, and make the assumption that the whole number lies at its middle point, we are representing it by a rectangle instead of by a trapezium with the side nearer the centre of the group longer than the other; a little consideration will show that that makes the second moment too great. Mr. Sheppard has shown that under certain circumstances it will be sufficient correction to subtract the fraction $\frac{1}{12}$ from the second moment calculated on the assumption of uniform distribution at the middle points of the groups to obtain a moment in a true approximation. On page 21 the corrected moment, m_2, is 6·564, while the uncorrected moment, μ_2, is 6·647. The correction will be $\frac{1}{12}$ only, if the difference between successive groups is one unit of abscissa; if the difference was h, we should have to multiply $\frac{1}{12}$ by h^2; but for practical work it is best to take the unit as the distance between groups which we are dealing with, and hence the correction $\frac{1}{12}$ is in the form which is of practical use.

The "standard deviation" is defined as the square root of the second moment about the centre of gravity. Professor Karl Pearson used σ to denote it, and σ is 2·562 inches in this case. It is sometimes more convenient to deal with the square root of twice the moment, which is called the modulus, and denoted by the letter c. Professor Edgeworth uses the modulus, whereas Professor Karl Pearson uses the deviation. We shall see the appropriateness of the modulus when we deal with the curve of error. The modulus for this group is 3·623 inches. It is a very remarkable fact that the modulus for the height of groups of men is almost universally very nearly 3·6 inches. Professor Edgworth gives a list of 10 such groups in the Jubilee volume of the *Journal of the Royal Statistical Society*: the moduli are 3·6 (United Kingdom), 3·6 (England), 3·4 (Scotland), 3·6, 3·7, 3·8 (United States), 3·7 (Belgium),

3·7 (Italy). I merely call attention to that in passing, to give an idea that the modulus is of real significance and not a mere arithmetical calculation.

Average Deviation.

For the next few minutes I propose to assume that the curve I am dealing with is symmetrical about its centre of gravity. The curve of heights which is sketched on page 21 is in fact very nearly symmetrical. If the curve is actually symmetrical all the odd moments are easily seen to be zero, while the even moments are not. Then this quantity σ or c, whichever we adopt, serves to measure the distance of the curve from its average, to use a clumsy phrase, or the dispersion about the average. Before discussing the appropriateness of this measurement I have to explain two simpler methods of measuring the same thing. One based on the first power of the deviations, and the other based on the distance between the quartiles. First for the average or mean deviation which, in the notation I am using, is called η. If we write down the deviations in the method just defined and add them up, we obtain zero; but if we treat all the deviations as positive and add up their absolute values we do not obtain zero. The calculation is as follows :—Treat the negative deviations and the positive deviations separately. The sum of the negative deviations is

$$\Sigma y(10{\cdot}0419-x)=10{\cdot}0419\times 824-6331,$$

from the figures to the left in the table on page 21. The sum of the positive deviations is

$$\Sigma y(x-10{\cdot}0419)=13100-10{\cdot}0419\times 1111,$$

from the numbers in the right compartment. The average deviation is, therefore,

$$\{13100-10{\cdot}0419(1111-824)-6331\}\div 1935=2{\cdot}01 \text{ (inches)}.$$

Probable Error.

The other simple method is based on the quartiles. Calculate the quartiles of this group by any of the methods already given, and you will find them to be approximately 65·8 inches and 69·3 inches. Since the median as given on page 21 is 67·566, one quartile is 1·78 inches below the median, and the other 1·72 inches above it. Half the distance between the quartiles is called the probable error. It is a term which is so firmly in use that there is no hope of

improving it, but it is one of the most erroneous terms in use in mathematics. Half the distance is 1·75 inches. If we take a person at random from this group and measure his height, it is as likely as not the height will be found to be between the quartiles, for the space contained between the ordinates at the quartiles is exactly half the whole curve, hence the phrase "probable error."

If we were dealing with the special distribution determined by the equation to the curve of error (see p. 34), we should have the following relations: average deviation = modulus $\div \sqrt{\pi}$, and probable error = modulus × ·4769. These relations are approximately true for this distribution of heights, for the values of the mean deviation found from these equations when the modulus is 3·623 inches are 2·04 and 1·73 inches respectively, while the numbers found above are 2·01 and 1·75.

These methods of describing groups are, however, applicable to groups which do not conform, even approximately, to the law of error. I shall now treat them without the assumption that they do conform. The probable error is the measure of dispersion, which is most quickly calculated. We can write down the quartiles very rapidly, and take half their difference at once. But that only takes into account the positions of the two quartiles, and does not take into account the positions of the extremes, but only their size, and, depending as it does only on two quantities, is liable to a large amount of accidental error. The mean deviation on the other hand takes into account the position as well as the number of all the quantities, and is therefore less liable to accidental error, and also it does not take at all long to calculate with simple numbers. The modulus and the standard deviation, again, take into account every observation, but they give extra weight to those which are a great distance from the average. In some cases that is right; in others it is not. If we are basing arguments as to the group and the shape of the group on probability, then very likely it will be correct to give this extra weight to an object which is far from the average, for the farther from the average the less the probability, and in some cases the probability diminishes very rapidly as we move from the average. If we are not going to make assumptions about the shape of the curve, nor apply the principles of probability, I do not know that we shall find any justification

for taking the square, rather than the mean, deviation. As a rough rule we may say that we pass appropriately from the probable error to the mean error, and from the mean error to the standard error as the curves with which we are dealing become more definite and perfectly continuous, and approximate more and more nearly to a curve with a definite algebraic equation. For very rough measurements which are not continuous and which are not to be corrected, the probable error, measured as half the distance between the quartiles, will very likely be the best measurement. As the curve attains a definite shape, and as we are able to treat the observations as more and more continuous, it will be well to take the mean error, and finally, if we have a perfect algebraic curve, then very likely it will be most correct to take the "standard deviation."

Measurement of Skewness.

Now to pass on to unsymmetrical curves. We have obtained by one of the averages the position of the curve and by one of these measures of dispersion one measure of its shape. We shall now obtain the measure of its want of symmetry, or briefly, of its skewness. Most curves have some degree of skewness; but in some cases it is negligible.

As an example of a curve with considerable skewness, we may take Diagram III, on p. 6. The curve is elongated to the right; the mode is to the left, the centre of gravity to the right of the median. This is the general order of these three averages. If a skew curve is formed by stretching a symmetrical curve to the right, the stretching shifts the centre of gravity, relatively to the median; or, from another point of view, if a curve is heaped up to the left and stretched to the right, experiment will show that the line through the median is to the right of the highest point.

There are very many possible ways of measuring this skewness. One obvious measurement is simply the distance of the centre of gravity from the median. Another is to use the quartiles. Call the positions of the quartiles, Q_1, Q_2, the position of the median, O, of the mode, M, and of the centre of gravity, G. In a symmetrical curve the distance Q_2O is equal to the distance OQ_1, whereas in a skew curve it will not be. In a skew curve stretching to the right, the upper quartile to the right is further from the median than the lower

quartile, and the difference between these two measures will form another means of estimating its skewness. The third method is to take the first power of the deviations, and compare the excess on one side the centre of gravity with the defect on the other. The fourth method is to take the third power of the deviations and consider its absolute magnitude. All these methods have their uses. I propose to deal with three of them. I will first take that which is arithmetically the simplest. The simplest measurement, that which you can calculate almost instantly, is the difference between the distances from the quartiles to the median. But that gives a concrete quantity; in the case before us so many inches; whereas it is convenient to measure the skewness as an absolute quantity, on a scale from $+1$ to -1; and we must therefore reduce this concrete quantity to an absolute quantity. The proper method of doing that is to divide it by the modulus, which is a concrete quantity, in this case so many inches. The one divided by the other gives an absolute measurement, which would serve to measure the skewness. But it is better to multiply that measurement by the constant 3·29 (see p. 36, below) before using it, to bring it into conformity with the theory of probability; in the same sort of way as the multiplication of the second moment by 2 to get the modulus brings the standard deviation into conformity with the methods of probability.

Another and yet simpler method of measuring almost exactly the same quantity, is to divide the difference between those two quantities by their sum, that is to say by twice the probable error; then if we multiply that by 3·14 (see p. 36, below), we shall obtain the same measurement very nearly as before. This method supplies a good rough measurement which is very rapidly calculated; we write down the median and the two quartiles, calculating them roughly or by one of the more complete methods given above, and at once write down the probable error; by this means the skewness of the group can be calculated in five minutes. But this measurement depends on the positions of three points only, which are subject to accidental errors, and the parts outside the quartiles have not much influence on the result.

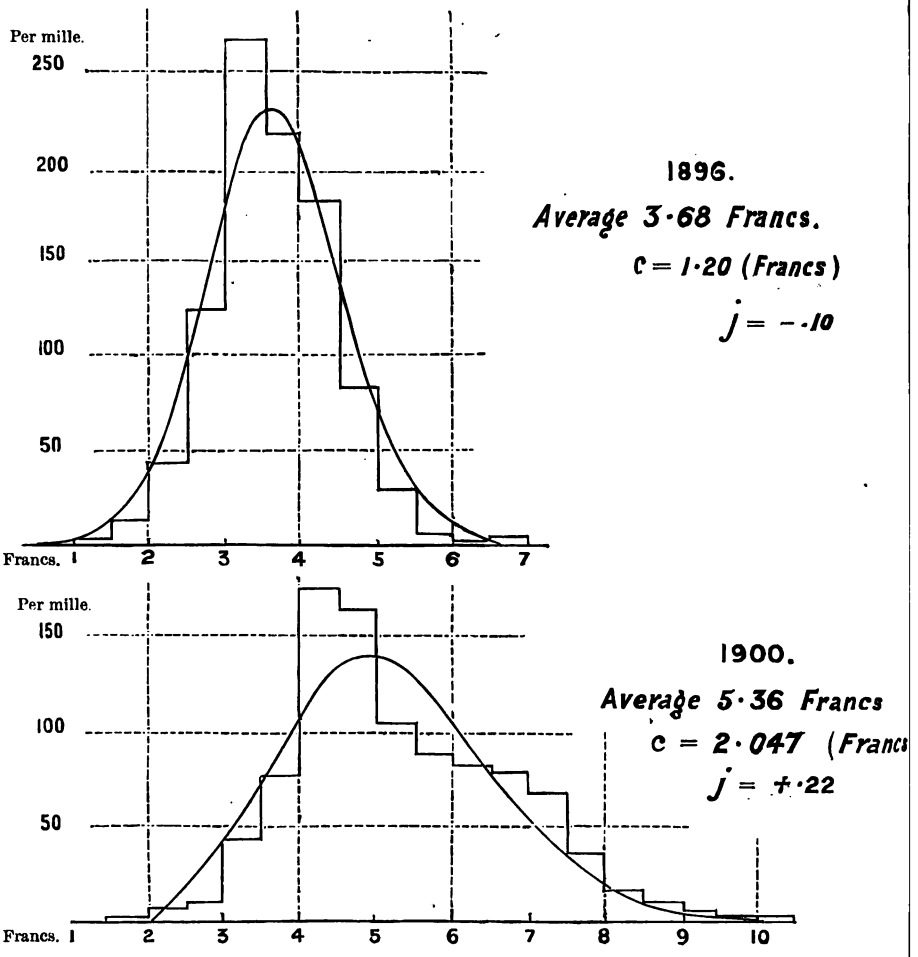
A measure, which is influenced by all the items, is obtained by taking the third moment about the centre of gravity; this in itself is a measure of the skewness, but it is not of the

right dimensions, for it is a concrete quantity of the order of a cube, as the deviations have been cubed ; to reduce it to an absolute quantity it must be divided by c^3. Calling this measure j, we have $j = \mu_3 \div c^3$, which in the group given on p. 21 is equal to +·016.

In a curve which is nearly symmetrical and approximates to the curve of error, the distance between the arithmetic average and the median will be $\frac{1}{3}$ jc, and the distance between the arithmetic average and the mode will be jc, and these relations supply a third method of estimating the skewness.

DIAGRAM VIII.

Daily Wages of Belgian Coal-miners.



First estimate the modulus, and then calculate the position of either mode or median and the arithmetic average, divide the distance by c or by $\frac{1}{3}c$, and we obtain j. But that is not an accurate method, if we use the mode, which cannot be precisely determined; while if we use the median, we are depending upon a single position. The formulæ to be preferred are $j = \frac{OQ_2 - OQ_1}{OQ_1 + OQ_2} \times 3{\cdot}14$, and $j = \mu_3 \div c^3$, the former perhaps when the curve is not approximately the curve of error.

The adjoining Diagrams illustrate the practical use of the technical quantities which I have now discussed. In 1896 the Belgian Government undertook an Industrial Census, and, amongst other things, they collected figures of the wages of most of the workpeople of Belgium. We have here in graphic form the daily wages of the Belgian coal miners in 1896. A supplementary enquiry was conducted in 1900 over nearly the same area, and the result is given just below. The methods we have developed give us a rapid means of comparing the results of those two enquiries. It is the rectangular figures only with which we have to deal at present. The average increased from 3·68 francs to 5·36 francs between the dates; the modulus from 1·20 to 2·047 francs, the skewness changed from a negative skewness of −·10 to a positive one of ·22. Those three statements rightly understood and interpreted give in a brief form the result of the Census. The average has increased, more money went in wages, and the modulus and standard deviation has increased very much. There was a development, therefore, of wages away from the average, either by highly skilled workers increasing their wages greatly, or by a body of unskilled workers coming into existence. If you look at the curve you will see the dispersion is chiefly increased to the right, and that increased standard deviation is due either to the inclusion of a higher grade of workmen than had been included before, or to the fact that the higher grades of work had obtained a great increase of wages. I am inclined to think it possible that the increase of dispersion is partly due to the erroneous inclusion of people in the second enquiry which were not included in the first, but I have no means of going behind the figures. The change of j comes from the same sort of reason, that a body of skilled workmen were

obtaining higher wages, or that the number of skilled workmen had increased. Either of these means would increase j in a positive direction. This use of the letters may be left for consideration.

Returning for a moment to the use of deviations in connection with the median and arithmetic average, I have to point out the curious relation between the two. The arithmetic average is that quantity from which the sum of the deviations is nothing, and the sum of the squares of the deviations the least possible. The second result is obtained instantly from the formula already given, $\mu_2=\mu_2'-\bar{x}^2$. The sum of the squares of the deviations from the arithmetic average is μ_2; the sum of the squares from some other origin is μ_2'; and from that formula μ_2 is always less than μ_2'. The median on the other hand makes the sum of the first powers of the deviations a minimum, and the sum of the zero powers zero. If we take the zero power of the deviations, each deviation is replaced simply by 1, and then from the definition of the median we find the sum of the zero powers measured from the median is zero. That the sum of the first powers is a minimum can be readily demonstrated, most easily by an analogy. Suppose that it is required to run from a telephone exchange separate wires to everyone of n places in a straight line, where should the exchange be placed, so as to use the least total amount of wire? At the median position. For if you move from the median position to the right or to the left you will find immediately that you are adding more wire than you are subtracting. Supposing there are 20 stations, and you have a position between the 10th and 11th; if you move to a position between the 11th and 12th, you have to increase your distance from 10 stations and diminish it from 9, in every case by the same length of the wire. The wires correspond to the deviations; and the sum of lengths of the wires is the sum of the lengths of the deviations. Consideration of this illustration will show that the sum of the deviations is a minimum when they are measured from the median, but that the median is not quite determinate, for if there are an even number of stations the sums of the deviations measured from all points between the two central stations are the same.

MEASUREMENT OF GROUPS.

THIRD LECTURE.

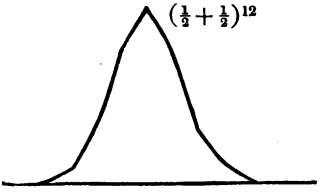
The Curve of Error.

THE subject discussed in this section is full of technical difficulties, and it will be impossible to cover the subject adequately in the short space allotted to it. It must then be regarded as containing rather a summary of those important points connected with the theory of error, which I shall have to use subsequently. While making it as complete as possible in itself, in several cases I shall have to ask acceptance without proof of results which I shall find it necessary to use at a future date.

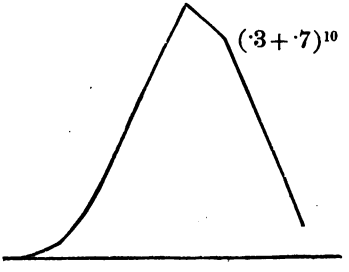
Among the various shapes assumed by groups of observations of any kind which are (as in the groups already taken) grouped in a more or less regular way about the central line, there is one distribution of the various deviations about their centre which is regarded as normal, and the curve representing it is called the curve of error. And it is the deduction of the equation of that distribution which I have first to deal with. After we have the equation we will discuss to what extent the normal curve is actually found in the kind of statistics with which we deal. The normal curve can be obtained from the statistics found in games of chance, or from the statistics which may be obtained by counting the occurrence of specified digits in mathematical tables, or from anthropometric measurements, or again from some groups of social statistics and from some groups of vital statistics. The deduction of the equation I am going to take is the only one which I think lends itself to purely algebraic treatment. Other deductions depend upon the use of differential calculus or even of the theory of functions.

Let us consider some occurrence for which the chance is p, the chance against q, so that $p+q=1$. Let us suppose

that the event which may or may not give the occurrence takes place n times again and again, and that in each n times we count how often success is obtained. For instance, suppose we pitch a coin n times and count how many heads are found and then repeat the n-fold experiment again and again and register in each case the number of heads, that would give a series of the kind I have in mind. For a small number of experiments, if each set of experiments contained 16 tries or any small finite number, it is easy to set down the probabilities of the various numbers of successes. And it is also clear as soon as the algebra of the method is tackled, that there is a limit towards which these chances tend as the number of experiments in each group is indefinitely increased. What we have to do first is to find the limit towards which such a series of experiments tends when the n is increased indefinitely.



The diagram annexed represents the various chances of the numbers of heads in the experiments of pitching a coin 12 times. The most probable number of heads is of course six, the least probable none, or 12, and the probability of 0, 1, 2, up to six, is continually increasing. If we erect 13 ordinates representing the probability of no heads, one head, and so on up to 12 heads, we get the diagram marked $(\frac{1}{2}+\frac{1}{2})^{12}$. If we take another kind of experiment where the chances for success and failure are not equal, *e.g.*, where the chance of success is ·3, and perform the experiment 10 times, we get the probabilities of one, two, and so on up to 10 successes represented by the following diagram :—



The first curve is of course symmetrical, the second curve unsymmetrical. What we have to do is to deduce the shape of the curve when the index is infinite, whether the chance in favour is one-half, or whether the chances for and against are unequal.

If p is the probability of an event, and $p+q=1$, then the probability of m successes in n trials is $p^m q^{n-m} \frac{\lfloor n}{\lfloor m \lfloor n-m}$, and successive values of m give the terms of the binomial expansion $(p+q)^n$.

Assume that np is integral. Let $np=r$, $nq=s$, $r+s=n$.

Denote successive terms by $u_0, u_1 \ldots u_n$.

Then u_s, which is the greatest term, $= \frac{\lfloor n}{\lfloor r \lfloor s} p^r q^s$.

$$u_{s+x}=u_s \frac{\left(1-\frac{1}{r}\right)\left(1-\frac{2}{r}\right) \ldots \text{to } \overline{x-1} \text{ factors}}{\left(1+\frac{1}{s}\right)\left(1+\frac{2}{s}\right) \ldots x \text{ factors}}$$

$$\log u_{s+x}=\log u_s+\log\left(1-\frac{1}{r}\right)+\log\left(1-\frac{2}{r}\right)+\ldots$$
$$+\log\left(1-\frac{x-1}{r}\right)-\log\left(1+\frac{1}{s}\right)-\log\left(1+\frac{2}{s}\right)-\ldots$$
$$-\log\left(1+\frac{x}{s}\right)$$
$$=\log u_s-\frac{1+2+\ldots+\overline{x-1}}{r}-\frac{1^2+2^2+\ldots+\overline{x-1}^2}{2r^2}$$
$$-\frac{1+2+\ldots+x}{s}+\frac{1^2+2^2+\ldots+x^2}{2s^2} \text{ \&c.}$$
$$=\log u_s-\frac{x(x-1)}{2r}-\frac{x(x+1)}{2s}-\frac{(x-1)x(2x-1)}{12r^2}$$
$$+\frac{(x+1)x(2x+1)}{12s^2} \text{ \&c.}$$
$$=\log u_s-\frac{x^2(r+s)}{2rs}+\frac{x(s-r)}{2rs}+\frac{x^3(r^2-s^2)}{6r^2s^2} \text{ \&c.}$$
$$=\log u_s-\frac{x^2}{2pqn}+\frac{x(q-p)}{2pqn}+\frac{x^3(p^2-q^2)}{6p^2q^2n^2} \text{ \&c.}$$

Let $x^2=z^2 \times 2pqn=z^2c^2$

$$\log u_{s+x}=\log u_s-z^2-\frac{z(p-q)}{\sqrt{2pqn}}+\frac{2z^3(p-q)}{3\sqrt{2pqn}} \text{ \&c.}$$

The assumption that np is integral made above does not affect the limiting form of the equation.

It is at this point necessary to consider which terms are to be rejected, when n is made infinite. If x is finite, if we move through only a finite number of terms from the greatest ordinate, the ordinate u_{s+x} equals the ordinate u_s. This part of the curve approximates to a horizontal straight line. To take a numerical instance, the chance of obtaining 499 heads in 1,000 tosses is practically equal to that of obtaining 500 heads. On the other hand if x is infinite, it appears that u_{s+x} is zero. If the figure is drawn so as to show finite values of x we obtain a horizontal straight line; but if an attempt is made to include infinite values of x, the curve becomes the axis of x and a finite vertical line through the origin.

But it becomes clear, if we examine the shape for different finite values of n, that the curve has a definite shape and finite curvature near the centre. Before we go further let us take an analogy. If we take an hyperbola and try to include the whole curve in our figure the curve will coincide with its asymptotes. In order to draw the curve so that the part between the asymptotes and the vertex can be seen, we must adopt a particular scale so as to obtain the length from the vertex to the centre as a finite quantity. Again, if we pass from the ellipse to the parabola by the process of pushing the centre to infinity you have, in order to obtain the finite part of the parabola at all, to make the hypothesis that $y^2|x$ is finite. In order to get the finite part of the curve of error we shall have to select that part where the ratio of x^2 to n is finite. Then it will be found that we shall obtain the part of the curve that has a definite curvature and a definite shape in a finite form. Let us assume, then, that $\frac{x^2}{n}$ is finite; and let us substitute for $\frac{x^2}{n}$ the quantity z^2 with the factor $2pq$. The reason for that factor will soon be obvious. Take $c^2=2pqn$, so that $x=zc$. We then obtain the equation $\log u_{s+x}=\log u_s-z^2$, when all vanishing terms are neglected. If the above deduction is carefully examined it will be found that all the terms omitted are infinitesimal in comparison with those retained, when n is infinite.

Removing logarithms, and writing y for u_{s+x}, we have

$$y=u_s e^{-z^2}=u_s . e^{-\frac{x^2}{c^2}}=u_s . e^{-\frac{x^2}{2pqn}}.$$

We are still at liberty to choose a scale for the ordinates, and it is most convenient to choose that which makes the greatest ordinate $=\frac{1}{c\sqrt{\pi}}$, for then the area bounded by the curve, and the axis of x becomes unity; then each part of the area represents the probability of certain occurrences, for the whole curve represents 1, which stands for certainty. An alternative is to take the ordinate as $\frac{N}{c\sqrt{\pi}}$, so that the area of the curve is N, where N is the number of experiments. Then the area standing on any part of the axis represents the most probable number of events corresponding to that part.

Now let us go back and take the terms we have so far rejected, which involve $1 \div \sqrt{n}$. Each of these contains the factor, $\frac{p-q}{\sqrt{2pqn}}$. It is convenient to call that quantity $2j$, for then we shall find that j has the meaning already assigned to it (see p. 21, and for proof see p. 36).

Re-writing the equation with that notation, and then expanding the part which contains j and neglecting the powers of j, we have

$$y=\frac{1}{c\sqrt{\pi}}e^{-z^2-2j\left(z-\frac{2}{3}z^3\right)}=\frac{1}{c\sqrt{\pi}}\cdot e^{-\frac{x^2}{c^2}}\left\{1-2j\left(\frac{x}{c}-\frac{2}{3}\cdot\frac{x^3}{c^3}\right)\right\}.$$

It is easily seen that $jc=\frac{1}{2}(p-q)=p-\frac{1}{2}=\frac{1}{2}-q$. The centre of gravity of that curve can be shown to be at the origin by integration. The area of the curve is of course the integral of ydx, taken between plus infinity and minus infinity. The part of the integral which does not contain j is a well-known definite integral, which equals unity. It can be seen that the part containing only odd powers of x does not affect the definite integral. Hence the area is unity.

Now let us calculate the error of mean square of the curve from the equation. It is obtained by multiplying the element of area $y.dx$ by its distance (x) from the centre of gravity, and adding up all the parts so obtained, and then dividing by the whole area, *i.e.*, unity.

It is easily seen that the j term does not enter into the result, which is therefore $\int_{-\infty}^{+\infty} yx^2.dx=\frac{1}{2}c^2$, by integration by parts. Comparing this with p. 21, we see that c, thus

calculated, is the modulus, as there defined. The third moment is $\int_{-\infty}^{+\infty} y.x^3.dx$, divided by the area, which is unity; integrating by parts we obtain that the skewness, as defined on p. 28, is equal to the j in this equation. The constants in the equation to the curve of error, as written above, are then the modulus and skewness as defined for curves in general. The average deviation, as defined on p. 24, is found by integrating $\int_0^{\infty} y.x.dx + \int_0^{-\infty} y.x.dx$, to be $\frac{c}{\sqrt{\pi}}$, and does not involve j.

The equation in its integral form is, if $Y\left(\frac{x}{c}\right)$ stands for area on abscissa from origin to x

$$Y\left(\frac{x}{c}\right) = \frac{1}{c\sqrt{\pi}}\left\{\int_0^x e^{-\frac{x^2}{c^2}}.dx \mp \frac{jc}{3}\left[1-e^{-\frac{x^2}{c^2}}\cdot\left(1-2\frac{x^2}{c^2}\right)\right]\right\},$$

the lower sign being taken when x is negative.

This does not admit of any simple evaluation, but it has been tabulated for a wide range of values of x*. From these tables it is found that the "probable error" for the symmetrical curve (where j is zero) is $c \times .4769$, which is written ρc. For the unsymmetrical curve the distances between the median and the quartiles can be shown to be $\rho c \pm \frac{2}{3} j\rho^2 c$†, while the distance between the centre of gravity and mode is jc†, and between the centre of gravity and the median is $\frac{1}{3}.jc$†, as used on pp. 27-29 above, where the resulting numerical values are given.

The effect on the curve of the j term is to stretch the curve to the right, heaping it on the left at the same time, the sort of figure which is indicated in the second diagram on p. 32. Actual examples of the curve for different values of c and j are given on pp. 21, 28.

The tables give the integral for the argument $\frac{x}{c}$, not for x, and before they can be used the observations must be

* See *Burgess's Mathematical Tables; Merriman's Least Squares*, p. 186; *Bowley's Elements of Statistics*, p. 281, and p. 332 (2nd Edition); and *Journal of the Royal Statistical Society*.

† See *Elements of Statistics*, p. 331. Hence, $OQ_2 - OQ_1 = \frac{4}{3} j\rho^2 c$, which gives results on p. 27 and p. 29.

reduced to the centre of gravity as origin and c as unit. Then if we find in the table that the integral function, *e.g.*, is ·455 when the argument $= +1{\cdot}387$*, we are to understand that ·455 of the whole area stands on the axis of x between 0 and 1·387 of the modulus. The tabular statement then shows the various fractions of the whole observations which may be expected (in an infinite number of experiments) to lie between the most probable value and various values with an assigned deviation from the centre. Thus with the symmetrical curve of error, one-quarter of the observations may be expected to be above the most probable value by not more than ·47 of the modulus, one-third by not more than ·68 of the modulus; all but 2 per 1000 are separated by less than 2·2 of the modulus from the most probable value; the chance of a deviation of 5 times the modulus is less than 1 in a billion.

Supposing we are given a set of observations which we have reason to suppose should arise from the distribution defined by the symmetrical curve of error, what particular curve of error are we to fit to our observations? The problem is not very important in itself, but the method of solution is very similar to the method which underlies the principle of least squares and of several other formulæ. The only things which we have a possibility of choosing are the abscissa of the centre of gravity and the modulus.

Let $x_1, x_2 \ldots x_n$ be the deviations of the observations measured from their average. The separate chances that these should arise if the equation of distribution is

$$y = \frac{1}{c\sqrt{\pi}} e^{-\frac{(x-k)^2}{c^2}} \quad \text{are} \quad \frac{1}{c\sqrt{\pi}} e^{-\frac{(x_r-k)^2}{c^2}}$$

where r is given successive values $1, 2 \ldots n$.

The chance that should occur together in a given group is, by multiplication,

$$c^{-n} . \pi^{-\frac{n}{2}} . exp. - \{\Sigma_1^n (x_r - k)^2 \div c^2\} = \mathrm{P} \text{ (say)}.$$

Now on what principle are we to find out the values of c and k? Of all the curves of error from which these observations may be supposed to have arisen there is one curve from which they would arise with the least improbability; to find this we have to make P a maximum. k and c are quite independent. Then the differentials of P with regard to k

* Which is the case when $j = +\cdot 073$.

and c must each be zero. The first gives that k is zero* and the centre of gravity of the observations is the origin. The second shows that $\frac{c}{\sqrt{2}}$ is the mean square of the x's.† So that to choose the normal curve which fits the observations best, in the sense that they would have arisen from that distribution with the least improbability, we must take for the centre of the curve the centre of gravity of the observation, and for the modulus the error of the mean square multiplied by $\sqrt{2}$.

It will be noticed in the proof that in a sense there is only one symmetrical curve of error. We can reduce any curve to the form $y = e^{-x^2}$, by suitable choice of scales for the co-ordinates; but if we are taking two groups measured in the same unit, for instance, both in inches, or shillings, or years, then the x axis has concrete units, the unit distance stands at one inch, one shilling, one year. And if we take two separate curves both measured in inches, work with the same unit of abscissa, and make the areas each unity, we do not get the same maximum ordinate. The finite part of the curve with the lower maximum ordinate stretches further to the right and left than the corresponding part of the other. As long as we deal with concrete quantities we shall find that the quantity c enters into the shape of the curve; and the comparison of any two curves is made by means of the values of c given in terms of the unit of abscissa. The quantity j is independent of all concrete quantities, and is an absolute measure of skewness, as already pointed out.

* $c^2 . \frac{\delta P}{\delta K} = (+2\Sigma x_1 - 2nk) . P = -2nk . P = 0$ when k is 0.

† $\frac{\delta P}{\delta c} = \left(-\frac{n}{c} + \frac{2\Sigma(x-k)^2}{c^3} \right) . P = 0$, when $c^2 = 2\Sigma x^2 \div n$, since k is 0.

$j = \cdot 06$

Unit 3·623 inches				Per 1,000			Difference from normal curve
Inches	τ	F(τ)	Y(τ)	Calculated	Actual	Difference	
				0	1	1	1
59	−2·357	·500	·511	0	3	+ 3	1
60	2·081	·498	·511	2	2	0	− 1
61	1·805	·495	·509	10	8	− 2	− 3
62	1·529	·484	·499	22	20	− 2	− 2
63	1·253	·462	·471	46	39	− 7	− 7
64	·977	·416	·431	81	79	− 2	+ 2
65	·700	·339	·350	119	125	+ 6	+12
66	·424	·226	·231	147	149	+ 2	+ 7
67	− ·149	·084	·084	155	166	+11	+11
68	+ ·127	·071	·071	138	127	−11	−17
69	·403	·215	·209	113	110	− 3	− 7
70	·679	·332	·322	73	79	+ 6	0
71	·955	·411	·395	47	47	0	− 1
72	1·231	·459	·442	25	28	+ 3	+ 4
73	1·507	·483	·467	12	13	+ 1	+ 2
74	1·783	·494	·479	6	5	− 1	+ 1
75	2·059	·498	·485	3	1	− 2	− 1
76	2·335	·500	·488	1	0	− 1	0
77	2·611	·500	·489			64	80

We will now use the height-statistics given on p. 21 as an example of the method of comparing a set of observations with the curve of error. In the first place we take the centre of gravity as the origin, namely:—67·54 inches. The modulus, by the method of moments is 3·623 inches, which is therefore to be taken as the unit. Thus 59 inches is 8·542 inches below the average, that is, 2·357 times the modulus. The latter number is entered under τ in the second column. All the others are calculated in the same way. Then turning to the

tables and finding what integral corresponds to the assigned values of τ, in the symmetrical curve of error, we write them under the heading $F(\tau)$. So we have that between the average and 59 inches ·5 of the whole curve is obtained, that is to say, one-half; in the next line, between average and 60 inches ·498 of the curve is obtained, and so on all the way down to between the average and 76 inches, when again half the curve is obtained, correct to the third decimal place. We should not get the the true half till we have gone to infinity, but the area of the curve beyond does not amount to one per mille of the whole. In this curve, for example, ·462 is the probability that the height of a person chosen at random lies between 67·54 inches and 63 inches, for ·462 is opposite 63 inches. The fraction of the curve is the same as the probability of the occurrence between the point given and the average.

The next column, called $Y(\tau)$, is obtained in a similar way from tables including the term involving j; the value of j is taken to be +·06 for reasons given below. The column following under "calculated" consists of the differences of the $Y(\tau)$ column multiplied by 1,000; the numbers so obtained are the numbers to be expected approximately between 59 and 60 inches, 60 and 61 inches, &c. The following column "actual" gives the actual occurrences per 1,000 in the same limits. The following column gives the differences in the various groups between the calculated and actual numbers. The greatest divergence is near the centre, where there are 12 more than were calculated. In the last column are given the differences if I had taken the normal curve instead of the skew curve. It is seen that by taking the curve as a skew curve the sum of these differences is diminished from 80 per 1,000 to 64 per 1,000.

I have now a rather difficult point to take with reference to one of those columns. Theoretically, j is calculated by the method of moments, the error of mean cube; but in practice that does not give good results. A single observation a long way from the average has a very great effect on the mean cube. So that if in this number of 1,935 persons we had included two persons from a nationality where stature was very low, or where it was very high, we should have instances at a long way along the group which would not properly vitiate the comparison of the curve of error, but would have a

very unfortunate effect upon the mean cube. Instead of having a homogeneous group, we should have a group of 1,933 people from one group and 2 persons from another group which would not belong to the same curve. There has been a great deal of discussion as to what should be done with such abnormal cases. A good way out of the difficulty is not to calculate j by the above method at all, but to calculate it by an *à posteriori* method, to choose that value of j which makes the misfit least. We have already chosen c so as to make the improbability less. Let us choose j by some similar test. The method I have adopted here is due partly to Professor Karl Pearson, and partly to Professor Edgeworth. It is to obtain figures (not given here) in such a form that it can be seen what value of j will make the sum of the absolute differences least. The value which satisfies this condition is found to be $j = \cdot 06$.* The value obtained from the moments method is ·016. This might have been used and would have given a result slightly better than the value $j = 0$. But I am inclined to say it is better to calculate j from the *à posteriori* method; I think it is quite as logical, and you are bound to get a better fit.

Professor Karl Pearson has given a test by which you can consider the following problem:—Supposing you had a population with certain characteristics, such as height, distributed according to a curve with a particular formula, required the probability that an assigned distribution would be obtained from the supposed distribution. Putting it into a more concrete way, suppose the equation of the height group for the whole population was this equation with $c = 3 \cdot 623$ inches, and $j = \cdot 06$: required the probability that 1,935 persons taken at random from the population would have the heights actually registered. Professor Karl Pearson has given a table† with the necessary figures for determining that probability. Calculation from his table on this distribution shows that if we take the symmetrical curve the probability of obtaining such a selection is ·4; that is to say, the chances are two in five that the 1,935 persons would not be further from the supposed distribution than they actually are. If we take the skew curve with $j = \cdot 06$, the probability is ·7; that is to say, the odds are seven to three that we should

* See *Journal of the Royal Statistical Society*, June 1902, pp. 337-8.

† See *London, Edin. and Dublin Phil. Mag.*, July 1900, p. 175.

obtain 1,935 persons as nearly conforming to this group as we have found. It is very difficult to argue back from the height of a person to the expression $(p+q)^n$, and I shall not at present attempt it. I have shown above that we should obtain this formula of the curve of error if we were dealing with chances, with events whose occurrence was, by those terms, in the binomial theorem. But the same equation will be obtained on very many other suppositions, and I have only taken the simplest. Before giving these, however, it is necessary to define a "frequency curve."

If we are dealing with a group of measurements which are distributed about their average so that the number of them which lie at any defined distance from their average, say between x and $(x+dx)$ in excess of it, can be represented by a definite function, say $f(x)$, of that distance, then the curve which represents this function, *i.e.*, $y=f(x)$, is the frequency curve of that group. If the unit of ordinate is so chosen that the whole area contained between the curve, the ordinates and its extremities, and the axis of x, is unity, then $\int_b^a y dx = 1$ if a and b are the limiting values of x; in many cases a and b are $\pm\infty$. Then if the quantity is selected at random from the group, the probability that it will lie between x_1 and x_2 is $\int_{x_2}^{x_1} y.dx$; the probability that it will lie between x and $x+dx$ is $y.dx$.

If we take the experiment I instanced at the beginning, the tossing of a coin, and make the number of times tossed very great, the chance of obtaining given deviations would be given by the curve of error, as already shown. This is the frequency curve for the group of experiments. Events are ruled by very different laws of distribution. We may have a very skew curve, as, for instance, in the curves of ages of wives in Yorkshire where the mode was a long way to the left of the average; the smooth curve which best fits those observations would be the curve of frequency for the ages of such persons. That is to say, if we draw this curve, representing as nearly as possible the observed facts, and we make this area equal 1, the area standing on the part of the axis between the 35 and 40-year marks would represent the chance of a person taken at random being between

35 and 40 years old. If we were given the age of a man who had a wife in Yorkshire and we did not know her age, that area would represent the chance that her age would be between 35 and 40. The life curve, to take another example, is a frequency-curve. To any frequency-curve we can assign a modulus calculated from the second moment. That tells one distinct fact as to the distribution about the average. The curve may have the greater part of its area to the left or to the right of the average, and it may have an asymptote as in the case of the curve of error; but there is in general only a small fraction of the area beyond two or three times the modulus, which may therefore be taken as indicating the practical extent of the curve. It is often useful to speak of the precision (h), instead of the modulus (c), where $h=\frac{1}{c}$. The greater h is, the more precise are the predictions that can be made as to a magnitude taken at random.

If we are dealing with frequency-curves whose practical range is small and whose modulus is finite, and if we take a great number of these frequency-curves, or rather if we have to select from a great number of things whose sizes are ruled by different frequency-curves, for example, if we make up a line of a great number of pieces of metal taken from different heaps with different frequency-curves for each heap, it is possible to find the frequency-curve for the sum of these elements, that is for the length of the line you have made. I will put that in different form with a different illustration. Suppose we are going to take 100 books, and we can select them from 100 different groups of books whose thicknesses are bounded within definite ranges and have a different modulus which can be assigned, required the breadth of 100 books put together. The most probable breadth will be that obtained by adding the averages of the 100 different groups. From the terms of the question it is obviously very improbable we shall get all the 100 below the averages of their respective groups or all above. The actual breadth will have a frequency-curve of its own about an average which is the sum of the averages of the groups from which you select. Its modulus can be shown to be the square root of the sum of the squares of the moduli of the original frequency-curves. Thus, to take a special case, if we are going to select two

things only which obey normal curves with the same modulus, the modulus for the sum is $\sqrt{2}$ times the modulus of either. The developments from this theory are of great practical importance.

If we take one sample at random from each of a number of these frequency-curves whose moduli are not very unequal, so that no one curve predominates, and add together the quantities so obtained, then the quantity obtained obeys the curve of error itself, whether the original frequency-curves were curves of error or not. I cannot give the proof here; the theorem as I state it is partly due to Laplace and partly due to Professor Edgeworth.* That is one of the most general statements of the cases in which the curve of error will arise; and that conception may properly be applied to the conception of height and the causes which determine the persons' height. No single cause has very great influence compared with others, so far as we know, and they all presumably have measurable effects whose frequency-curves are definite. Thus, we might expect *à priori* the frequency-curve of heights to be the curve of error.

Another illustration is supplied by the grouping of school children in a particular grade.† I took one of the most populous grades in the Report of the St. Louis Public Schools, U.S.A., grouped the children according to their ages, and fitted the curve of error by one of the methods I have described. The curve of error with $c = 1{\cdot}68$, $j = {\cdot}073$, fits the observations closely. If we think of the causes which determine the position of a child in a particular grade or class, I think we shall find that they are akin to those I have supposed in my statement as to causes which lead to the asymmetrical curve of error. But it would be absurd to go back and try to re-value p, q and n, the quantities on which the algebraic proof of the equation depended. We could find out, of course, what chances would produce this particular distribution; but they would have no necessary relation to the facts. The idea I wish to give is that we can obtain the equation of the curve of error in the form I am using it on a very simple supposition; and it can be obtained from many other suppositions which cannot be given in lecture work.

* *See* Edgeworth, in *London, Edin. and Dublin Phil. Mag.*, 1892, p. 429.

† For the numbers and diagram, *see* Elements of Statistics, 2nd Edition, Appendix.

MEASUREMENT OF GROUPS.

FOURTH LECTURE.

The Method of Least Squares.

Suppose that we make a great many measurements of the same quantity by several different methods; and that, as is generally the case, the measurements differ from each other, owing to imperfections of instruments, or by the numerous accidental circumstances that attend any involved observations. Let us assume that the measurements which could be made by the first method are grouped according to the frequency-curve $y=\frac{1}{c_1\sqrt{\pi}}e^{-\frac{x^2}{c_1^2}}$, those by the second method according to $y=\frac{1}{c_2\sqrt{\pi}}e^{-\frac{x^2}{c_2^2}}$, and so on, a definite normal curve for each method. Suppose we make n measurements, one of each kind. It is required to find what is the most probable value of the distance to be measured. All the x's we are dealing with are errors in our measurement. From the series of partly erroneous measurements it is required to find the most probable value. That is the problem it is attempted to solve by the method of least squares. As a second question, it is required to determine the precision of the result, that is, to state the probability that it is correct within assigned limits.

Before going further I must call attention to one very important point in the reasoning. In the reasoning on which the method of least squares is based it is assumed that the frequency-curves are normal curves of error, as written above.

If the frequency-curve is not a normal curve of error the method breaks down at the first step. That I shall have to return to later. With regard to the moduli, we may either suppose that we know them by some *à priori* method, as is sometimes the case; or that we know them by having made similar experiments at some other time, *e.g.*, if we are dealing with a group of height measurements where the modulus is three inches generally; or we may find them from the experiments themselves. A useful way is to repeat the measurement by each method, say, 100 times, and from the internal evidence find out what the moduli are. We assume that the moduli are fixed quantities, quantities which we cannot affect, and that they are known or previously determined quantities. What is the probability that a certain series of errors should result in n observations? Let $x_1, x_2, \ldots . x_n$ be the differences from the unknown true value which arise from n different methods taken in one series; what is the probability that those particular n deviations will occur at once? The probability is obtained by multiplying together the probabilities of their separate occurrences. The probability of the error x_1 occurring, when the modulus is c_1, is from its curve of frequency $\frac{1}{c_1\sqrt{\pi}} e^{-\frac{x_1^2}{c_1^2}}$. The probability that the n will all occur is obtained by multiplying n such quantities together, that is, $\frac{1}{\pi^{\frac{n}{2}} . c_1 c_2 \ldots . c_n} e^{-\Sigma\frac{x^2}{c^2}}$. Here the only variables are the x's. Now, that probability will be greatest when the index of e is greatest, that is when $\Sigma\frac{x^2}{c^2}$ is least. Thus, from all the possible values of the unknown true measurement, the system of errors which we have found would arise with the least improbability when $\Sigma\frac{x^2}{c^2}$ is made the least possible. That is the statement which is at the basis of the method of least squares. In the particular case, when we take all the the observations by the same method with the same curve of frequency, so that c is the same for all the observations, the minimal condition is satisfied when the sum of the x^2 is a minimum; and we have already seen that that sum is made least when the unknown value is taken to be the arithmetic average of the obtained values. Let me re-state this theorem

in other words. Suppose we start to measure a particular object by the same method again and again. Then, the measurements we obtain would come with the least improbability when the sum of the squares of the deviations is a minimum; and that condition is satisfied if we take the arithmetic average of our measurements to be the unknown true quantity. This statement is a particular case of the method of least squares.

When we have grasped that initial principle, the rest of the investigation is only a matter of the differential calculus; there is nothing special about it. We have to write down all the equations that connect the quantities we are measuring, and then by the ordinary processes of the differential calculus express the conditions that the sum of the squares of the errors shall be a minimum, and these will give enough equations to solve for all our unknowns. I will illustrate that algebraically by a particular case. Take the case with which we have already dealt, namely, that in which we had the ages of the wives in Yorkshire. There we obtained a somewhat irregular curve representing the numbers at different ages, and we smoothed that curve by putting parabolic curves of the fourth degree through various points; and it will be remembered that we had to change the constants in our equation according to the particular group of five points selected. Now let us assume that we have a parabolic equation of the third degree in this form,

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3.$$

This equation has four unknowns; we can therefore make it pass through any four assigned points, but we cannot make it pass through five assigned points. Suppose that we wish to determine an equation of the third degree which will pass near the five points, then we will apply the method of least squares to that problem. Let the co-ordinates of the actual observations be (x_1, m_1), (x_2, m_2), and so on. Let the corresponding points which we are to find on this particular curve be (x_1, y_1), (x_2, y_2), and so on. The point $(x_1 y_1)$ will be near, but probably not coincident with, the point $(x_1 m_1)$. The difference between m_1, the observation, and y_1, which would be given by the curve which we have not yet determined, is the error of the observation. We are to determine the constants so that the sum of the squares of those errors shall be least. Writing

that a little more fully, and substituting for y in terms of x, we have that

$$\Sigma_1^5(m_1 - a_0 - a_1x_1 - a_2x_1^2 - a_3x_1^3)^2$$

is to be a minimum.

In that expression the variables are the four a's, which have to be determined so as to make the expression a minimum. Therefore we must differentiate that expression, when it is written out, with respect to a_0, a_1, a_2, a_3, and equate these partial differential coefficients to zero, obtaining as many equations as we have unknowns. Then we have to solve the equations so obtained.

After a little simplification the following equations are obtained:

$$5.a_0 + \Sigma x_1.a_1 + \Sigma x_1^2.a_2 + \Sigma x_1^3.a_3 - \Sigma m = 0$$

$$\Sigma x_1.a_0 + \Sigma x_1^2.a_1 + \Sigma x_2^3.a_2 + \Sigma x_1^4.a_3 - \Sigma mx = 0$$

$$\Sigma x_1^2.a_0 + \Sigma x_1^3.a_1 + \Sigma x_3^4.a_2 + \Sigma x_1^5.a_3 - \Sigma mx^2 = 0$$

$$\Sigma x_1^3.a_0 + \Sigma x_1^4.a_1 + \Sigma x_1^5.a_2 + \Sigma x_1^6.a_3 - \Sigma mx^3 = 0.$$

The chief thing I want to say abut these equations is, that they are so complicated, and a solution is so laborious, that they must be put out of court for all ordinary calculations. If you wish to construct a new table which will be of some general use, it may be worth while to go through the solution, but not for any single practical piece of work. Every one of those separate terms, Σx_1, &c., have to be calculated arithmetically, and the equations have to be solved. Even in this simple case we have four equations each containing four functions. In Merriman's "Method of Least Squares," the simplest methods for that evaluation are given. Many terms drop out, and the evaluation is possible; and in some cases we can so choose our origin and take advantage of certain points of symmetry in the equations, that the work can be simplified. In this particular case a simple solution has been given by Professor Darwin.*

Fitting Formulæ to Observations.

Before we look for another way, let us consider again whether the assumptions on which the above method depend are justifiable, or will justify the great effort which would be

* See Darwin, "On Fallible Measures," *London, Edin. and Dublin Phil. Mag.*, July 1877; used in *Elements of Statistics*, pp. 256. 257.

necessary to solve the equations. I think it will be found that in general they do not. If we look back through the argument, it will be seen that the original assumption is that the difference between m, the actual number of persons observed, and y, the number obtained from the equation, belongs to the normal curve of frequency; and so in every case where the method of least squares applies we have an observed measurement, and we obtain a theoretical measurement, and we assume that the difference between the two belongs to a normal curve of frequency. Before we can make that assumption we must verify that the conditions, under which the normal curve of frequency is obtained, are satisfied. We are not in a position to do that, if we depend only on the algebraic proof given above, without investigating the deductions of the equation of the curve of error resting on other hypotheses. But to my mind there is no proof yet given which does show that the normal curve of error will be obeyed in the circumstances I have just mentioned; and Professor Karl Pearson has shown that in very many instances the normal curve is not obeyed. So the theory is at any rate difficult to establish *à priori*, and is not supported by universal experience. I think, with all the deference that is due to Professor Karl Pearson, that the matter yet wants more practical experience before it can be fully decided. It would be unsafe in the present state of the argument on the one hand to say that the normal curve of frequency may be expected; or on the other hand to say definitely that it is not to be expected, because it has not been universally found. That is too difficult to deal with at all thoroughly here. The reason I have gone so far into it is this: if the method of least squares is very difficult to apply, and if it is neither supported sufficiently by theory nor by experiment, then it seems expedient to try some other method. A purely empirical method would be this: Instead of making the sum of the squares of the deviations a minimum, make the sum of the first powers of the deviations, all reckoned as positive, a minimum, that is to say, remove the square outside the bracket in the expression on p. 48. But it is not at all easy to make that sum a minimum, because all the terms have to be taken as positive, and we do not know until we have finished our work which terms are naturally positive or which terms are negative. Professor Edgeworth has given a method of

getting the solution when there are only two unknowns.* When there are three unknowns I believe there is as yet no practical solution.

Another method, still taking the method of least squares as the basis, but avoiding the very complex solution, is to choose the coefficients, so that the curve will pass through exactly the four points assigned; and then re-calculate them, so that the curve shall exactly pass through four other assigned points; and so continually calculate again and again the coefficients, getting a series of curves. Then from the various values of the coefficients so found, choose those coefficients which appear to give the best results. It is really a makeshift method. I think it has been often employed, and the results have been very satisfactory. If, by one method or another, you get coefficients which make the theoretical curve pass near the original curve, it does not matter by what process you have got them. Such a method as that, I think, is in general use for approximating to the population in inter-censal years. I think the Census Office has never published this method; but as far as I can find out, the method employed is as follows: Supposing certain points represent the population at the various dates at which it is exactly enumerated, then if, as a first hypothesis, we assume that the population increases in geometric progression between two enumerations, we obtain a simple curve passing from one point to the next. Then assume again that from this Census to the next there is another increase in geometric progression, and we find that the two curves never have exactly the same constants. Then obtain some method for passing from one curve to the other without a sudden break of curvature, reject the parts of the curves near the Census years, and replace them by a curve which gradually passes from one to the other. That is a purely empirical method, and I think it is the one adopted. It is in some such way as this that we can go to work if the method of least squares is too complicated.

The third method, to which I wish to call attention very particularly, proceeds in quite a different way. We tabulate our observations as before, and write down the equation of a curve which is assumed to fit them, with unknown constants; calculate from the observations the moments—first, second,

* See Edgeworth, "On a New Method of Reducing Observations," *Phil. Mag*, 1888; used in *Journal of Royal Statistical Society*, June 1902, p. 341.

third, fourth (as many as there are unknowns)—about the centre of gravity, by the method used above, and calculate the moments from the assumed curve in terms of the unknowns. Equating the moments found from the observations with the moments found for the assumed curve, we have these equations determining the constants. For example we may take the instance already discussed, when we found a skew curve of error to fit certain observations. The general equation to the skew curve of error being given, by the help of the integral calculus we stated the values of the first, second, and third moments in terms of c and j; we equated these to the moments calculated from the observations, and thus found c and j. We need to calculate as many moments as there are unknowns in the particular equation selected. For instance, in Makeham's formula there are four unknowns, and we have to take four moments. In the normal curve of error there are two unknowns, its centre and the modulus; two moments are therefore sufficient to find the normal curve of error by this test. In the skew curve of error, the quantity j has to be determined in addition. In the empirical equations given by Professor Karl Pearson in his well-known paper on the measurement of skew groups, which was published in 1895 in the Proceedings of the Royal Society, there are four unknowns, and therefore in general he needed four moments. In the parabolic interpolations, such as I have used in these lectures, there are as many unknowns as we like to take. If we stop at x^3, we need four moments. In Professor Pareto's empirical equation for the grouping of the incomes of the people of a country there are two unknowns. The equation is as follows: $y = \frac{A}{x^a}$, where y is the number of persons in receipt of income x, and A, a are constant. It is also given in a developed form with one more constant. It is supposed that the index a is nearly the same for all countries, while A varies from country to country. You could obtain those values by the principle of least squares, or by equating moments. This is not the place to criticise the equation: I only give it as an example of algebraic equation for statistical grouping. We see then how to obtain sufficient equations for the unknown constants, and so we come naturally to the question of what is the justification for this method. I think I must refer you, in

general, to Professor Karl Pearson's paper for the justifications, because it is his method, and in particular he has quite recently published a paper in the journal *Biometrika*,* going very carefully into this whole method; and all I can do is to simply follow in his steps. The method depends on a purely empirical basis, not on any *à priori* theory. By its means we do, as a matter of fact, obtain an equation which fits the observations. But, incidentally, Professor Karl Pearson shows that the results obtained are, in general, the same as those obtained by the method of least squares. Without basing his system upon the coincidence at all, he does obtain the same results. The advantage of the method is, as he has also shown in the same paper, that the solution of the equations obtained is very much easier than the solution of equations obtained by the ordinary method of least squares. I hesitate to go further into this subject because it is Professor Karl Pearson's subject, and all his papers are very easily accessible. He has shown that empirical algebraic formulæ can be found for a very wide range of groups, and in every case he has fitted equations to the groups by the help of this number of moments. He has then found that the equations so obtained do fit the groups exceedingly well. Groups may, perhaps, contain 30, or 40, or 100 measurements, but the constants at disposal are only 4. If you calculate these 4 constants by any method and obtain, as a result, the equations which fit a wide range of observations, you have a strong empirical justification for the method. I believe that is the justification which Professor Karl Pearson gives for the method. But we are met face to face with this difficult question, which it is impossible to deal with here and now: How far ought we in such investigations to take empirical formulæ which are only justified by their results, and how far should we base our reasoning on *à priori* assumptions as to the nature of error, and as to its occurrence, assumptions which underlie the theory of probability, and from such assumptions obtain our equations? Should we obtain our equations with the view to fitting the result, or should we obtain our equations from *à priori* reasoning and see how far they fit the results? To my mind we have not nearly enough experience in the matter at present. We have not sufficiently tested the fitting of groups to the *à priori*

* *Biometrika*, April and December 1902.

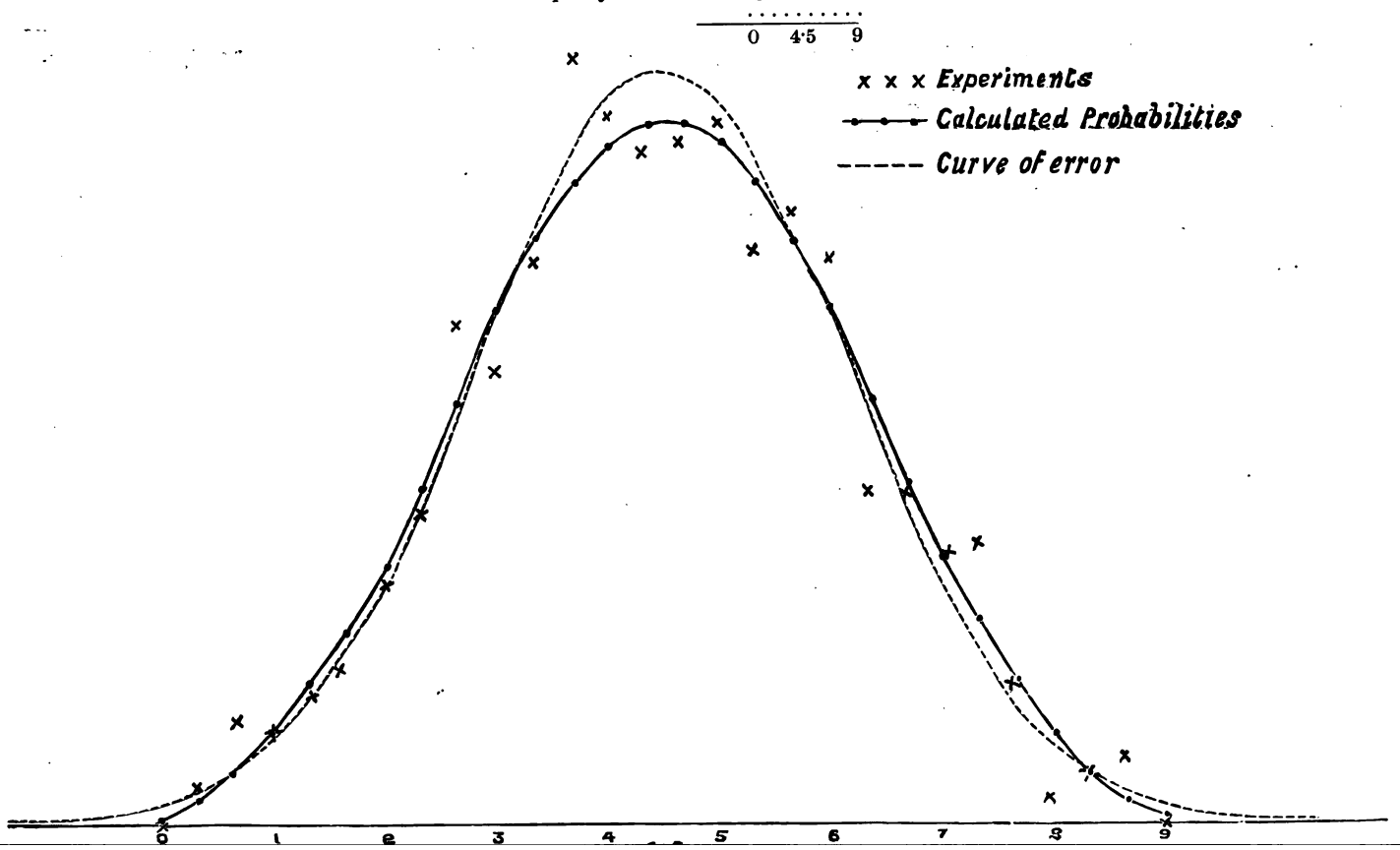
equations, nor have we yet sufficient experience to say that the empirical method is universally satisfactory because it has been found to fit wide ranges of groups. At that point I must leave the discussion.

Uses of the Curve of Error.

Whatever may be the ultimate decision in the questions which I have thus stated, there are certainly many uses for the curve of error in the form in which I gave it in the last lecture, quite independently of the discussion we have just been engaged in. In what I have been recently saying I have been following, as far as possible, Professor Karl Pearson's method. In what I shall say now I am following Professor Edgeworth's work. I do not mean that the two are contradictory in any way; I wish to indicate that I am trying to summarize the present position of this question on the lines of the two most eminent authorities in this particular work. For clearness, I repeat the method of generating the curve of error given on p. 43. Suppose we have a number of frequency-curves, each of small and limited range, that is to say, of great precision, its modulus being small; let the moduli of n such curves calculated from the squares of the deviations be $c_1, c_2, \ldots c_n$. The curves may be of any shape, except that no finite part of their areas may be at a great distance from their centres of gravity. Suppose we take a_1 observations belonging to the first curve, a_2 out of the second, and so on, and add them together; the curve of frequency for the resulting sum is the normal curve of error with modulus $\sqrt{(\Sigma a^2 c^2)}$. If instead of taking the sum, we take any other function to which the sum is the first approximation, the curve of frequency for the values of this function is likely to approximate to a normal curve of error; but we will here limit ourselves to the sum. The following diagram and the experiment on which it depends illustrate this theory. I took Chambers' mathematical tables, and chose three digits at random and took their average, and repeated this a thousand times. The curve of frequency of the 10 natural digits is a straight line; you are as likely to get any one of them as any other, if you select a suitable part of the tables. I have represented that curve of frequency by ten dots. It is limited at both ends, its modulus is fairly

Diagram IX.

Frequency line for the digits 0 . . . 9



big, viz.: 4·06, and it supplies a very severe test of the principle I have enunciated, because we have a curve of frequency which is absolutely different from the normal curve of error; it does not approximate to it in any way whatever. The actual probabilities of the occurrence of various numbers are the successive coefficients in the expansion of $(1+x+x^2+\ldots+x^9)^3 \div 1{,}000$. Comparing these with the result of the experiment we have the following table:—

Average of 3 digits taken at random	No. of times this Average	
	Was actually found	Might be expected every 1,000 times
0	0	1
$\frac{1}{3}$	4	3
$\frac{2}{3}$	11	6
1	10	10
$1\frac{1}{3}$	14	15
$1\frac{2}{3}$	17	21
2	26	28
$2\frac{1}{3}$	33	36
$2\frac{2}{3}$	53	45
3	48	55
$3\frac{1}{3}$	60	63
$3\frac{2}{3}$	82	69
4	76	73
$4\frac{1}{3}$	72	75
$4\frac{2}{3}$	73	75
5	75	73
$5\frac{1}{3}$	61	69
$5\frac{2}{3}$	65	63
6	60	55
$6\frac{1}{3}$	35	45
$6\frac{2}{3}$	35	36
7	29	28
$7\frac{1}{3}$	30	21
$7\frac{2}{3}$	15	15
8	3	10
$8\frac{1}{3}$	6	6
$8\frac{2}{3}$	7	3
9	0	1

It is not my point here to show that those figures are what you would expect to get; what I wish to show is, first, that the successive probabilities, when they are plotted out, resemble the curve of error; and, secondly, that the experiment tends to fit a normal curve of error. In Diagram IX the continuous line with dots on it is the frequency which you would expect. The broken line is the curve of error, with the

same area and modulus, and the crosses are the positions obtained from the actual experiment. It is seen that though we started with a frequency curve which was a straight line, that the theoretical curve which we obtained for the average of only three terms selected from it is already so much like a curve of error that you would mistake it for one, if a model was not traced on the paper; and that the actual experiment supports the same view.

We note that the modulus calculated from the squared deviations for the natural digits is 4·06, and that from the formula $\sqrt{(\Sigma a^2 c^2)}$ given above the modulus for the sum of three digits should be $\sqrt{3 \times (4{\cdot}06)^2} = 7{\cdot}032$, and for the average of three digits should therefore be 2·344. The modulus of the curve given by the calculated probabilities of the various numbers is 2·345, while that calculated from the results of the experiment is 2·358. The averages are 4·5 (theoretical) and 4·494 (experimental).*

Construction of a Group from Samples.

The theory which I have just enunciated, for the proof of which see the reference given on page 44, is, that if we start with any frequency-curves, and take our examples from them, one from each or many from one, and take the average, we shall obtain a curve which becomes more and more like the curve of error as we extend the number of our examples, and as the frequency-curves satisfy more and more nearly the limited conditions which are laid down for them. Now, that is not only a mathematical theory: it has very great practical importance. Supposing that we take a number of samples out of a large group, how near the true average may we expect to get? If the curve of frequency of the group was a curve of error, we can at once write down the probability of different divergencies. If we have a curve of error with modulus c, and we select n samples at random from it, and then take their average, the modulus for their sum is from the formula already given, $\sqrt{nc^2}$, and hence that for their average is $\frac{c}{\sqrt{n}}$. The precision of the arithmetic average varies inversely as the square root of the number of items, a very well-known principle. I wish to show how this theory can be adapted to

* See also *Edgeworth*, in Jubilee Volume of the *Journal of the Royal Statistical Society*, p. 186.

curves of frequency other than the normal curve of error. Suppose the original curve of frequency to be any curve whatever, a curve of survivors for example, I do not assume any particular shape to it. Suppose we go through an experiment, taking, we will say, m examples at random from it, and repeat the process k times. [In the experiment just discussed m was only three, and k was 1000.] Though the original numbers do not obey the normal curve of error, yet the average of m of them may be expected to, when m is sufficiently great. Let c be the modulus for the group of averages of m samples; then $\frac{c}{\sqrt{k}}$ may be expected to be the modulus for the average of the whole mass of km samples. Thus, in the above experiment, c was 2·35, k 1000, and $\frac{c}{\sqrt{k}} = \cdot 064$; the known average for all digits, which formed the original curve of frequency is 4·5, the average for the 3,000 selected, in 1,000 groups of three, was 4·494; the difference is one-tenth of the modulus just calculated; so small a difference might be expected once in nine trials.

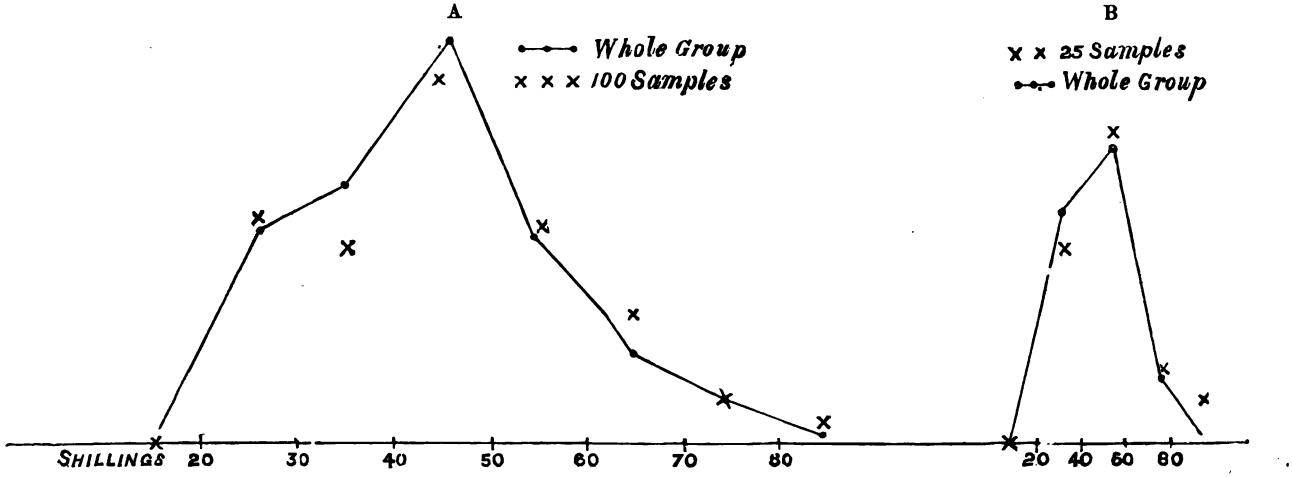
Thus, whether the curve of frequency of the original group is the normal curve of error or not, the precision of the average of a great number of samples is proportional to the square root of that number.

Now let us see how to construct not merely an average, but a whole group, by the method of samples.

Gazette Prices of Wheat per quarter.

Price	(1) No. of Cases	(2) Frequency per 100	(3) 100 Samples	(4) Frequency per 25	25 Samples
Under 20/-	3	0	0	0	0
20/- to 30/-	111	18	19	10	8
30/- „ 40/-	134	21	16		
40/- „ 50/-	206	32	30	12	13
50/- „ 60/-	105	17	18		
60/- „ 70/-	47	7	11	3	3
70/- „ 80/-	26	4	4		
Above 80/-	4	1	2	0	1
	636	100	100	25	25
Average	43/9	...	45/4	...	46/6

DIAGRAM X.


A
Whole Group
100 Samples
B
25 Samples
Whole Group
SHILLINGS
20
30
40
50
60
70
80
20
40
60
80

The table and diagram give the result of an experiment in such construction. The material of the experiment is of no importance here; I merely took the most accessible figures to conduct the experiment, namely, the official Gazette prices of wheat for the 636 months for which they are recorded in the statistical abstracts, and regarded that as a group of things which I was going to build up by sample. For complete illustration I had to take a group I knew, and then to take samples of it. In general, of course, the group is not known, but has to be constructed from the samples. The actual group is that given in Diagram X in the continuous lines. To obtain the samples, I took Chambers' mathematical tables, and assigned to particular numbers, from 001 to 636, certain months, and took 100 numbers of three digits at random. Next, I wrote down the prices in the 100 months corresponding to those 100 numbers, and grouping them in 10*s.* groups, obtained the numbers given in the third column above, and also given by the crosses in the Diagram X (A). I next selected 25 samples by taking the first 25 of the 100, and I grouped the figures in 20*s.* groups, and obtained the numbers given in the fifth column and by the crosses in Diagram X (B). What rule have we for deciding how near the true group the sample is? In the third division, for instance, between 30*s.* and 40*s.* in the whole group, there are 134 instances, and 21 per cent. of the area is between 30*s.* and 40*s.* If we take 100 things at random out of the whole group, how many of that 21 per cent. are we likely to get? This is a simple problem in probability: if n samples are taken, the chances that 0, 1, 2 . . . n will come from a given part, which is to the whole as is p to 1, are the successive coefficients of the expansion of $(q+p)^n$, where $q=1-p$; as n increases we approximate to a curve of frequency with modulus $\sqrt{2pqn}$ (see p. 34). In the third division $p=\cdot21$, while n, the whole number of samples in the first experiment, is 100. Here $\sqrt{2pqn}=\sqrt{(2\times\cdot21\times\cdot79\times100)}=5\cdot8$. The difference between the actual number per 100 in the group, namely, 21, and the number found in the sample, namely, 16, is less than the modulus. In all the other cases in both experiments the differences are within the "probable error" (which is ·47 of the modulus, see p. 36). We have thus found a criterion of the divergencies to be expected between the distribution of

magnitudes in a group of samples and the distribution in the unknown group from which they arise.

As regards the precision of the averages of the samples, the modulus for the original group is about 19*s.*, and, therefore, the moduli for the averages of 100 and of 25 samples, respectively, are $19s. \div \sqrt{100} = 1s.\ 11d.$, and $19s. \div \sqrt{25} = 3s.\ 10d.$ The averages found from the samples are actually 45*s.* 4*d.* and 46*s.* 6*d.* which are, respectively, 1*s.* 7*d.* and 2*s.* 9*d.* in excess of the average of the whole group.

The experiment, therefore, forms a good illustration of the theory, and on consideration it will, I think, be found that the theory is in strict accordance with common-sense and common experience.

MEASUREMENT OF GROUPS.

FIFTH LECTURE.

Correlation between Two Groups.

LET there be n pairs of measurements (x_1y_1) (x_2y_2) and so on up to (x_ny_n), the members of each pair having some determinate connection with each other; for example, suppose that the x's are the ages of the wives in the group taken above, and the y's the ages of their husbands, x_r and y_r being the ages of a married couple. This is the example discussed below. Or suppose that x_r is the age at which a man dies, and y_r the age at which his father died; or suppose that x_r, y_r are measurements of physical characteristics of the same man. Or again, x_r might be a death rate, in a year in which y_r was the average temperature. It is required to measure the relationship between x's and y's so as to answer this question: Given one of the x's, assign the probable value of the corresponding y. For example, given the age at which a man died, assign the most probable age to which his son will live. Or, taking one member of the group of wives at random, state the probabilities of the age of her husband. We have in fact to give numerical expression to such statements as these: A high death rate goes with a low temperature; a long-lived father has long-lived sons; for two statements where two measureable quantities are connected in that way, where in common parlance we connect them with simple adjectives, we have to find a numerical or mathematical expression for the relationship. First suppose that there is no *causal*

connection between two groups. Then if we select any particular place on the axis on which the x's are measured, and mark in the corresponding y's, we shall get a group of y's whose average is equally likely to be above or below the average of all the y's. Suppose we choose a group of wives, between the ages 25 and 30, mark in the ages of their husbands, and mark the average of such ages, if there is no connection between the ages of the one group and the ages of the other, the average of the group so taken will be near or equal to the average of the whole group of husbands, namely, 42 years. And so, if we take another period and mark in the various ages of the husbands we should again find the average near the average of the whole group. If the x's are represented on a horizontal axis, and the y's are measured vertically by points placed above the values of x which are their pairs, then if there is no causal connection between the magnitude of the x's and of the y's, the averages of groups of the y's corresponding to assigned intervals on the axis of x will all lie near the horizontal line through the averages of the y's. They will not lie on it, but the best straight line we can draw near these points will be a horizontal line through the average; that is obvious as soon as the statement is understood.

But now suppose there is a causal connection between the two sets of measurements; suppose, for example, that a high value of x goes with a high value of y. Then if we start from the average value of x, which we may assume for the moment corresponds to the average value of y, and pass to the right and choose a group at a place above the average for the x's, the y's which are obtained for that group will be distributed about an average above the line. And as we continually mark off the averages for group after group by points, they will lie on some curve which tends upward to the right from the origin and downwards to the left. (See for example Diagram XI.) If, on the other hand, a high value of x went with a low value of y, there is a change of sign; the series of averages would go down to the right and up to the left. The exact method of drawing a line through these points I do not propose to discuss very minutely. We could draw a smooth line by the methods discussed in the first lecture, or a freehand curve. We can either draw a straight line as near as possible to the dots, or we can draw a curve. I shall

not discuss the general shape of that curve; I shall merely assume that, from the observation or otherwise, we can draw that curve. And since in any series of observations the particular averages are liable to slight displacement, in a finite number of observations we do not get the most probable point with each average, and must smooth the line in the way we have discussed. We may assume an equation, $y=f(x)$, which gives the average of the y's for the particular values of x; that is only giving a general form to the statement, that a value of y is connected with a value of x by a determinate equation.

This equation, of course, only gives the position of the averages of the selected groups of y's. Everyone of these groups has its own frequency-curve. If we select again the ages of the husbands of those wives whose ages are between 25 and 30, we can draw a frequency-curve for that group of husbands, but the centre of that frequency-curve will no longer be at the average age of all the husbands, if there is causal connection between the groups; but as the group taken is below the average of the wives, the centre of this curve will be below the average for the husbands. It is not necessary, in general, to make any attempt to draw this frequency-curve point by point, but only to take its centre and in some cases its modulus. Instead of dealing with arithmetic averages, we may equally well use the medians of the groups.

We might take, for example, such a question as this, a very old question: Has the price of wheat anything to do with the marriage rate? In such a case as that we plot out the prices of wheat in different months or years along the axis of x, and put in ordinates showing the average marriage rate when the wheat was that particular price, and the direction of this line or the form of this curve would give, within certain limits dealt with below, the answer to this question, whether there was a connection between the two or not. If we do obtain from our observations that there is a tendency upwards to the right and downwards to the left, or *vice versa*, we have found that there is something common in the system of causation which produces the two sets of phenomena. We cannot say that the x's are the cause of the y's, nor *vice versa*, but only that the two phenomena are not absolutely independent.

We have to find a numerical measure of that dependence. If the curve that we obtain is a straight line, we have only to find a means of calculating its inclination. Before proceeding to this, let us spend a few words on the case when the curve is not a straight line. Suppose that we have sufficient observations to determine by experiment and observation the actual shape of this curve from large groups, we could, without applying any further theory whatever, establish the connection between the x's and the y's; the curve can be plotted out, and given algebraic expression, if possible; and then we should be able to say that for a particular value of x the most probable value of y was the one obtained on this curve. We could have a curve simply from experience, and use the experience with similar phenomena at another time. For instance, if we had that experience of the length of the lives of the children of parents who lived to various ages, we should be able from this empirical curve, to say if a man's father lived to a certain age then the chances of the life of the son are given by a frequency-curve whose centre was found from the empirical diagram, and whose shape might very likely be known also. In many cases, however, the curve of averages is approximately a straight line. Even if the approximation is not very exact, it may be useful to calculate the inclination of the straight line that passes nearest the averages. Let us suppose that we have the equation of this line, $y=ax+b$. Consider any observation x_r, y_r; if this observation lay exactly on that line, y_r would be ax_r+b. If the observation does not lie on the line, its distance from it, measured parallel to the axis of y, is $y_r-(ax_r+b)$. To obtain the best values for a and b, which are the only unknown quantities, we can proceed* by the method of least squares, and make the sum of the squares of such quantities as $y_r-(ax_r+b)$ a minimum. Then the differentials of $\Sigma(y_r-ax_r-b)^2=u$ (say) with regard to both a and b must be zero.

Thus
$$\frac{\delta u}{\delta a}=2a\Sigma x^2-2\Sigma xy+2b\Sigma x=0,$$
$$\frac{\delta u}{\delta b}=2nb+2a\Sigma x-2\Sigma y=0.$$

* See below p. 73.

Choose the axes so that both the x's and the y's are measured from their averages, then $\Sigma x = 0 = \Sigma y$, and the equations give us $b=0$ and $a=\frac{\Sigma xy}{\Sigma x^2}$; the line required passes through the origin, and its equation is $y=\frac{\Sigma xy}{\Sigma x^2}\cdot x$. Let σ_1, σ_2 be the standard deviations of the groups of x's and of y's, so that $n\sigma_1^2=\Sigma x^2$, $n\sigma_2^2=\Sigma y^2$, and let $r=\frac{\Sigma xy}{n\sigma_1\sigma_2}$; then the above equation becomes—

$$y=\frac{\Sigma xy}{n\sigma_1^2}\cdot x=\frac{r\sigma_2}{\sigma_1}x,$$

that is,

$$\frac{y}{\sigma_2}=r\cdot\frac{x}{\sigma_1}.*$$

In order to make r symmetrical, it has been necessary to divide by σ_1 and σ_2, that is, to measure x and y by their standard deviations. It is a very natural thing to do. Before we can get any numerical comparison, we must reduce them to some common measure, and a common unit which we can very reasonably adopt is the standard deviation for each of the two things. If we are dealing with the question I suggested just now—the marriage rate and the price of wheat—we cannot compare shillings with a rate per thousand, but we can compare a ratio of the number of shillings to a standard number of shillings, with the ratio of the rate per thousand to a standard rate per thousand. We are then comparing absolute instead of concrete quantities. We should get similar equations if we used the modulus instead of the standard deviation, or the probable errors, or the mean deviations. For rapid work we could replace the σ_1 and σ_2 by the probable errors, which are proportional to the standard deviations in curves which approximate to the curves of error. It is to be noticed that we can express the quantity r in the following form: r is the average of such products as $\frac{x_r}{\sigma_1}\cdot\frac{y_r}{\sigma_2}$. r is called the *coefficient of correlation*. It is not difficult to show by pure algebra that the quantity r so determined must lie between $+1$ and -1†; and that r equals $+1$, only if the ratio of every x to its corresponding y is

* The last few paragraphs are substantially the same as those given by Mr. Yule in the *Journal of the Royal Statistical Society*, 1897, p. 817 seq.

† See *Elements of Statistics*, p. 319.

identically the same as the ratio of every other x to its corresponding y, so that the ratio $y:x$ is constant, and equal to $\frac{\sigma_2}{\sigma_1}$. If the ratio is constant and $= -\frac{\sigma_2}{\sigma_1}$, r becomes -1, and an increase of x corresponds to a diminution in y. Thus r is always between $+1$ and -1, and between these limits there is a scale of correlation. For instance, we can say that the correlation between two sets of phenomena is ·6 or $-$·3. Of course, when one is first introduced to a new scale of any sort the numbers in the scale convey no meaning; it is a matter of experience to attach the right value to the different magnitudes in the scale. Perfect correlation can be understood from the statement that groups are perfectly correlated if a deviation of a member of one always equals the deviation from the average of the corresponding member of the other multiplied by an assigned constant. If the two things, marriage and wheat prices, were perfectly (negatively) correlated, you would be able to establish some such equation as this: An increase of ·1 in the marriage rate is always found with a diminution of 6*d.* in the price of wheat. Of course, such a rigid relation is never obtained unless there is some physical cause binding the two things together. As the ratio of corresponding pairs tends to constancy, the correlation becomes more and more perfect. That must be regarded as a definition of correlation.

Now consider the sum of the products of x and y, and let us write X for $\frac{x}{\sigma_1}$, and Y for $\frac{y}{\sigma_2}$.

If there were no correlation, if we selected the values of Y which corresponded with a particular small range of values of X, we should be likely to find a negative value to neutralize each positive value of Y, and the products arising from that range of X's would tend to zero, and the greater the number of terms the less the distance of their average from zero. But directly there is any bias towards getting the positive value of Y for this particular range of X's, as we increase the terms we may still get negative terms here and there, but on the whole we shall get positive terms; and so on, all the way up the scale of X's. When there is correlation it is clear that the sum of the products tends to be greater than where there is none. Thus it seems probable

from first principles that the quantity r thus calculated will make a good measure of correlation.

There is an important caution to be given in the use of this formula. If, from two series of phenomena which were absolutely unconnected, we took a limited number of examples, say a thousand, and worked out the value of r, we should not obtain exactly zero, or rather the chances are very much against obtaining exactly zero, even if there was no correlation; and if we took a very small number of examples the chances are very much against obtaining anything near zero. As we increase the number of samples, if there is no correlation, the coefficient will tend more and more nearly to zero. What we require before we can use the coefficient is some criterion to enable us to know whether the formula is significant, or whether the actual number might have arisen if there had been no correlation whatever. Such a criterion is given below on p. 88.

Ages of Husbands	Ages of Wives: 15 to 20	20 to 25	25 to 30	30 to 35	35 to 40	40 to 45	45 to 50	50 to 55	55 to 60	60 to 65	65 to 70	70 to 75	75 to 80	above 80	Wives: No.	Median Age
20—25	2	21	6	..	..	..	..	..	..	..	..	..	..	..	29	23·1 years
25—30	..	28	49	9	1	..	..	..	..	..	..	..	..	..	82	26·8 ,,
30—35	..	4	31	49	9	1	..	..	..	..	..	..	..	..	94	31·2 ,,
35—40	..	1	7	29	43	9	1	..	..	..	..	..	..	..	90	35·9 ,,
40—45	..	..	2	7	25	36	7	1	..	..	..	..	..	..	78	40·7 ,,
45—50	..	..	1	2	7	22	31	6	1	..	..	..	..	..	70	45·6 ,,
50—55	..	..	..	1	2	6	18	28	5	1	..	..	..	..	56	50·2 ,,
55—60	..	..	..	..	1	2	5	18	17	4	1	..	..	..	48	55·05 ,,
60—65	..	..	..	..	..	1	2	4	9	11	2	..	..	..	29	59·4 ,,
65—70	..	..	..	..	..	..	1	1	3	6	6	1	..	..	18	63·75 ,,
70—75	..	..	..	..	..	..	..	..	1	2	3	3	1	..	10	68·05 ,,
75—80	..	..	..	..	..	..	..	..	..	..	1	1	1	..	3	71·60 ,,
80—	..	..	..	..	..	..	..	..	..	..	..	..	..	..		76·17 ,,
Husbands No.	2	49	96	97	88	77	65	48	36	24	13	5	2		602	
Husbands Median age (Years)	22·8	25·8	29·3	34·0	38·9	43·9	48·8	53·7	58·6	63·2	68·0	72·5	76·4	80·2		

Average age of Husbands, 42·16 years. Standard deviation, 12·6 years.

Average age of Wives, 40·11 years. Standard deviation, 12·1 years.

Coefficient of correlation is ·96 approximately.

The numbers given are in every case the nearest thousands.

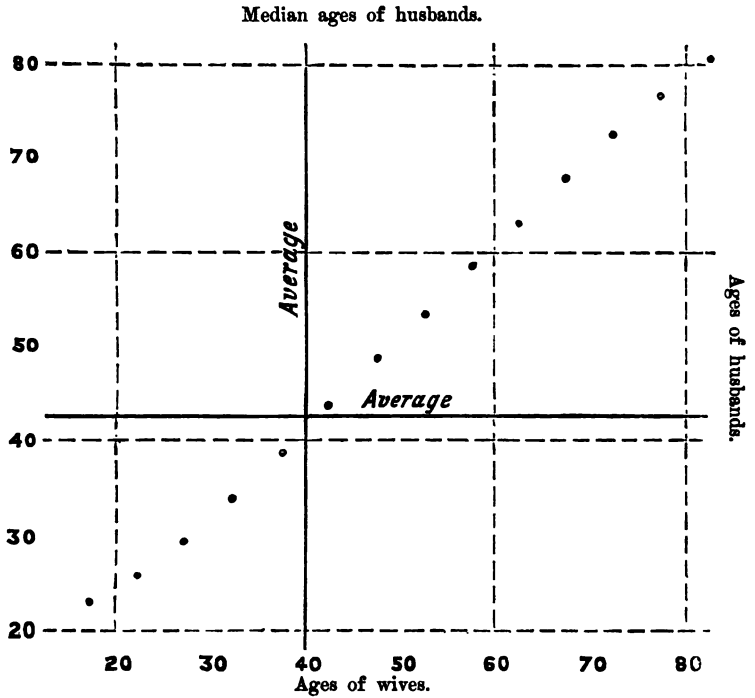
A numerical example I have prepared will put the calculation in a clearer light. The table here given shows the numbers (to the nearest thousand) of the wives and husbands in the County of York in 1901 at various ages,

grouped in periods of five years. For example, if you take the husbands' ages from 40 to 45 and look along the line you will see that there are two wives between 25 and 30, seven between 30 and 35, 25 between 35 and 40, and so on. It was not practicable to deal with 610,000 cases, and I have therefore dealt with the thousands only, and approximated throughout the calculation. The fact that the numbers run diagonally down the table as they do shows at once there is correlation. I have taken a case where the correlation is nearly perfect. If we had a table where the correlation was very small we should find the numbers distributed in random fashion all over the table. In such a list of figures as this it is not very practicable to take the arithmetic average. It is easier and as accurate to take the medians. I have approximated to the medians for all the groups, both horizontal and vertical, by the methods already explained. To take a particular example, consider again the husbands who are between 40 and 45 years of age. If you look along the list you will find there are in all 78, and that the median age is 40·7 years. Or if you take a vertical column, if you choose those wives who are between 40 and 45, and look vertically downwards, you will find that there are in all 77 of them, and that the median age of their husbands was 43·9.

The diagrams show the medians graphically. In the first the ages of wives are measured horizontally, those of husbands vertically. Above the middle point of each five-year period is placed a dot indicating the median age of husbands whose wives' ages come in that period. Thus, looking upwards from the position of 37½ years, the middle age of the group of wives between 35 and 40, you will find that the dot indicating the median age of the husbands is placed at the 38·9 years. You see that the points so obtained lie very nearly in a straight line. At the top and at the bottom the line becomes a little bit curved, for the influences of the lower and upper limits of ages make themselves felt. If we tried to get a normal distribution for the group of wives who are under 20 we should be getting husbands at 13 and 14 years of age, and at the other end of the scales we should have got husbands at ages at which there are no people alive. The fact that the scale is limited at both ends is the cause of the deflection of that curve from the straight line. We have now to measure the inclination of that line. In the case I have taken, where

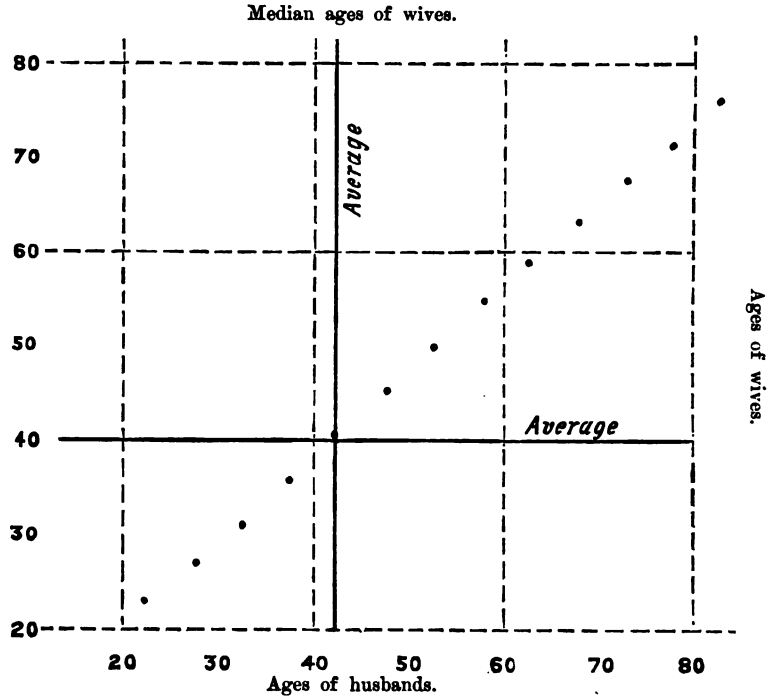
DIAGRAM XI.

Showing correlation between ages of husbands and their wives.



The tangent of the inclination of the line through the dots is ·97 (nearly).

DIAGRAM XII.



The tangent of the inclination of the line through the dots is ·92 (nearly).

the correlation is so perfect, there is no difficulty in measuring the line, because if a straight line is drawn through three or four of those points it passes very near the others. But in other cases, it is not so obvious which straight line is to be drawn; and then we can proceed by the method of least squares already taken, or you can proceed by the following practical method which yields good results:—Mark out two lines horizontally and vertically through the averages of the two groups, and rotate a ruler through their point of intersection until the same number of dots is found on the one side of it as on the other. It will be found that that method gives a definite position of the line which passes very near the points; it is a purely empirical way; but as the coefficient of correlation need generally not be calculated with great minuteness, it will in general be sufficiently correct. It is often absurd in cases of probability to work out the results with very great accuracy. The line is not drawn in the diagram above, because it would have obscured the dots; but underneath is given the tangent of the inclination to the horizontal of the line which would satisfy the conditions, the tangent of this angle is ·97. The second diagram is constructed in a similar way, for the median ages of wives, whose husbands are in a given group; the tangent of the inclination of the line through the points is now ·92. The average age of the husbands is 42·16 years, with standard deviation 12·6 years; the average age of the wives 40·11 years, with standard deviation 12·1 years. The statement we have now obtained is of this sort:—If we are dealing with a man whose age is h, in excess of the average, and we wish to know the age of his wife; the value of w in the equation $h-42{\cdot}16={\cdot}92(w-40{\cdot}11)$ is nearly the most probable value of her age. That comes at once from the geometry of the second diagram. From the first diagram we obtain similarly:—Given the age of a woman as being w, so that the deviation from the average is $w-40{\cdot}11$, then the median age of the husband group is $42{\cdot}24+{\cdot}97(w-40{\cdot}11)$. We shall probably also need to know the curve of frequency for each of these groups. Unless there is a reason to the contrary, I think in general that we may assume that the curve of frequency for a selected group is similar to the curve of frequency for the whole group from which it was selected. So that we can

calculate the standard of deviation for this particular curve of frequency when you know the standard of deviation for the whole group. That is to say, we can ascertain the chance that the age of the husband of a particular woman is any assigned number of years above or below the age here selected. In the little table given above we can actually find these small curves of frequency; for instance, in the ages of wives 30 to 35 years of age the curve of frequency for the husbands goes as follows:—9, 49, 29, 7, 2, 1.

The above is the graphic way of working out the question, We have now to show its relation to the formula for r, the coefficient of correlation. The quantity $r\sigma_2/\sigma_1$ in the equation $y = r\sigma_2 x/\sigma_1$ is the quantity evaluated by the diagram as ·97. that is the tangent of the inclination of the line to the axis of x, where ages of husbands and wives are measured on the axes of x and y respectively, and σ_1, σ_2 are the standard deviations for wives and husbands. If I had reduced all the measurements to the standard of deviation beforehand, r, the coefficient of correlation, would have been the tangent of the inclination of the line. The question which is the easier to work, decides which of the two methods you adopt. If you work it as I have done, with the same scale of years vertically and horizontally, you would have to say that $r = \cdot 97 \times \frac{\sigma_1}{\sigma_2}$, and from the lower diagram that $r = \cdot 92 \times \frac{\sigma_2}{\sigma_1}$; whence r is the geometrical mean between ·97 and ·92, between the tangents of the inclinations of the line calculated on the two different hypotheses, namely, ·945; and $\frac{\sigma_1}{\sigma_2} = \sqrt{\frac{\cdot 92}{\cdot 97}} = \cdot 974$.

Now let us proceed to calculate r by the formula $\frac{\Sigma xy}{n\sigma_1\sigma_2}$. It is, of course, a long business, and I shall not give the work completely; I shall only indicate the way in which it was done. The problem was to find that product for 610,000 pairs, which, of course, is a prohibitive piece of work, and cannot be done accurately, because the ages are not given except in 5 yearly limits. We proceed by approximation. First of all, I neglect all the numbers below 1,000; secondly, I assume that the numbers left are at the middle of their respective groups. Then I deal with the 60 or 70 numbers in the table on p. 67 in the following way. Select a group, *e.g.*, wives whose ages are

25 to 30; the middle, 27½, is 12·6 years below the average age of all wives; express this and other deviations in terms of the standard deviation, 12·12; 12·6÷12·12=1·04. That is the y term to be applied throughout this group of husbands. Go through a similar process for the x's. This 6 is at the middle of the group 20–25 years, namely, 22½, which is 19¾ years below the average age of all husbands, and that in terms of the standard deviations is about 1·6; work out the other deviations, which are in arithmetic progression, in the same way. Then multiply the numbers in the group, 6, 49, 31, 7, 2, 1 each by its deviation, add, and multiply the sum by the deviation 1·04 for the group.

EXAMPLE:

AGE OF WIVES, 25—30

DISTANCE FROM AVERAGE—1·04 OF STANDARD DEVIATION

Age of Husbands	Distance from Average	No.	Product
20–25	−1·6	6	− 9·6
25–30	−1·2	49	−58·8
30–35	− ·8	31	−24·8
35–40	− ·4	7	− 2·8
40–45	+ ·03	2	+ ·1
45–50	+ ·4	1	+ ·4
	(Of Standard Deviation)		Sum −96·5

Corresponding part of $\Sigma \frac{x}{\sigma_1} \cdot \frac{y}{\sigma_2}$ is −1·04 of −96·5 = +103·6

Some of the resulting terms will be negative unless the correlation is considerable. Add these terms, and divide the sum by n (in this case 602), and the coefficient of correlation ·96 is obtained. In the method I suggest using we do not deal with any large numbers at all. The number is not far from the geometric mean of ·945 found by the graphic method above. Also $\frac{\sigma_1}{\sigma_2} = ·96$ (instead of ·97 as reckoned above), $r \times \frac{\sigma_2}{\sigma_1} = ·92$ (the same as above), $r \times \frac{\sigma_2}{\sigma_1} = 1·0$ (instead of ·97).

JUSTIFICATION OF THE FORMULA FOR r.

In the method of finding the formula for r on page 64, we used the method of least squares without examining its suitability. I will now give reasons, which have not, so far

as I know, been previously offered, in favour of this method. If the figures we are dealing with belong to the normal curve of error, there is no difficulty. If their curve of frequency has any other form, still the averages of selected groups, represented by the dots in Diagrams XI and XII, are governed by normal curves of error (*see* page 56). Let $\bar{y}_1, \bar{y}_2, \ldots \bar{y}_m$, be the averages of groups of y's, containing respectively $k_1, k_2, \ldots k_m$ items, whose x values are $x_1, x_2, \ldots x_m$; so that the k_r x's which, in the grouping of x's adopted, are in a small group whose centre is x_r, have y-pairs whose average is $\bar{y}_r$. Let $y=ax+b$ be a line which contains the values of y from which the observed $\bar{y}_1$, $\bar{y}_2$, &c., are deviations. Then $(\bar{y}_r - ax_r - b)$ is a quantity whose frequency-curve is $\eta = \frac{1}{c\sqrt{\pi}} e^{-\frac{\xi^2}{c^2}}$, where c the modulus is inversely proportional to $\sqrt{k_r}$, k_r being the number in the x_r, $\bar{y}_r$ group. The probability of such deviations occurring together is (as on page 46) a maximum, when $\Sigma k_r (\bar{y} - ax_r - b)^2$ is a minimum. Equating the partial differentials of this sum with reference to a and b to zero, and remembering that $\Sigma k_r \bar{y}_r = 0 = \Sigma k_r x_r$, if the deviations are measured from the general average, we have, as on page 65, $b=0$, and $\Sigma k_r (ax_r^2 - x_r \bar{y}_r) = 0$. Hence, $a = \Sigma k_r \bar{y}_r . x_r \div \Sigma k_r x_r^2 = \Sigma xy \div n\sigma_1^2$, where the summation extends over all the pairs. Then, as before, $r = \frac{a\sigma_1}{\sigma_2} = \frac{\Sigma xy}{n\sigma_1\sigma_2}$.

Consideration of the nature of the formula will, I think, lead to the conclusion, that the coefficient of correlation calculated by the formula is a good measurement of correlation, whatever curves of frequency you are dealing with; and it is surprising how very rapidly a small extent of correlation makes itself felt, even when you deal with quite a few examples. If n is only 20, you will soon find whether there is correlation or not by this formula. If you select groups where there is no correlation the criterion, discussed below, shows that the correlation is not significant; but directly there is likely to be correlation between the groups, this formula for r shows it. The coefficient of correlation can be used then in a very large region of cases in which it is required to test the connection between two series of phenomena. In particular, it can be used to decide whether two series of phenomena are entirely unconnected or not, which subject necessitates a preliminary treatment of the nature of series.

MEASUREMENT OF SERIES.

SIXTH LECTURE.

Series.

I PROPOSE to deal in this lecture, first of all, with series in general, and then with the comparison of and correlation between two series. By a series I understand a list of numerical events recorded at regular intervals, for example, recorded once every year. In representing a series by a diagram we measure time on the horizontal axis, and dividing it up into years, we erect an ordinate at the point corresponding to each year, to represent on a suitable scale the magnitude at that particular year. The question whether we should represent these magnitudes by dots or lines or rectangles is important, but it is decided on the principles discussed when we were dealing with the representations of groups, and we need spend no more time on the analysis now. Perhaps the most natural way of representing such series is to erect a series of rectangles whose areas are proportional to the successive magnitudes; but if we leave the diagram in that form it will not be very clear, it will be very ugly, and certainly this is not a neat way of finishing the representation. The next step is to draw a continuous line to replace the rectangles; the commonest way of doing this is, to mark the middle points of the tops of the rectangles, and join those points by straight lines; but this method is erroneous, for the same reason that it was erroneous in the representation of a group. We need to draw a continuous line so that the areas

contained in the rectangles in the first place, and by the curved trapeziums in the second, shall be equal in every case; but this more correct curve is practically coincident with the erroneous straight lines; it makes very little difference in practice which of the two we draw; they give certainly the same optical impression. If, however, we do replace the rectangles by a continuous line we are making an assumption which is sometimes justified, but sometimes not; by the fact of drawing a continuous line we give the impression that the event represented is continually taking place. This is correct in representations of births, deaths, and marriages, and it is partly correct in representing imports and exports by curves but it is not correct in the representation of events which only occur once each year. These are details which are easily analysed.

Classification.

The series, or the curves which represent them, can be divided into three main classes: periodic curves, symptomatic curves, and others; or instead of "others," we may say curves with random fluctuations. Periodic curves are those where similar fluctuations recur at equal intervals of time, as the annual fluctuation of temperature recorded month by month. Symptomatic curves are those which have a definite tendency up or down, a "symptom," though short periods may obscure it, as the death rate since 1870. A curve, which is neither periodic not symptomatic, may often be regarded as having random fluctuations about a stationary average, as a curve representing the annual averages of any meteorological phenomena, such as average temperature year by year. In the Diagram XIII* all four curves are symptomatic; the first three are downwards, and the last upwards for the first 30 years and then nearly level. The series represented in Diagram XIV has apparently random fluctuations. These curves are not periodic in any strict sense.

Periodic Curves.

The first thing to discuss is, how to disentangle the period from the symptom when a periodic curve is also symptomatic, or how to measure the period if the curve is not symptomatic. There is not space to discuss the matter completely, and I want rather to indicate the methods, and leave their consideration

* See p. 81.

to the reader. A curve often suggests two things: first, that there is a regular period, and, secondly, that there is a movement apart from the period. Assume that we are dealing with monthly observations and an annual period. To obtain the movement apart from the period, take the averages of the 12 months of each year and mark them on the diagram; these points would show the average rate for the year, when the readings of the vertical scale have been adjusted. But there is something arbitrary in beginning the year at the 1st of January. The deaths, births, and marriages, and any other figures we deal with are probably independent of that particular beginning of the year, and if we make comparisons it may be better to take other periods to start with; for instance, the fiscal year begins on April the 5th. We want a continuous representation, which we can obtain as follows:—First take the average from January 1st to December 31st. Then the average from the 1st of February to January 31st, and so on until we get 12 dots every year. It is clear that the curve through these points cannot have any sudden fluctuations; the curve so obtained shows the symptom when the period is eliminated. The theory underlying this method is quite simple. If we take any particular 12 months, we shall include the whole influence of the period, the excess in one part and the defect in another, and if we average them we shall probably get the number which would have occurred if there had been no period, and if the flow had been regular. It is approximate only, because the various small fluctuations will affect the average, and it can be improved by smoothing the curve. If the series is not symptomatic the resulting smooth curve should be a horizontal straight line.

Now, in order to measure the period as apart from the symptom, the only method is to write down the rates for the 50 Januaries which we may be dealing with, and take the arithmetical average, the mode, or the median of these; to repeat the process with the Februaries, and so on; and then to represent the successive averages for the 12 months by a separate curve, which is best drawn with a base line through the general average of all the data. We thus get such a curve as that given by the graph of $y = \sin x$, from 0° to 360°. The justification of the method is simple. In the 50 Januaries we include one January from each part in the symptomatic

curve. All the excesses due to the symptomatic tendency will be counter-balanced by the defects, or will tend to be counter-balanced by the defects, due also to symptomatic tendency. They will only tend to be counter-balanced; for if we take the 50 Januaries we include among them some extraordinary months, and some months whose deviation from the annual average is quite small. The accuracy with which we may expect to get the true January reading is proportional to the square root of the number of times taken, from the theory of averages discussed above. In carrying out the method, we implicitly assume that the causes which decide the symptom and the causes which decide the period are independent, while generally they are not independent. If there is an increasing death rate or an increasing want of employment at the same time that the winter is especially severe the one will accentuate the other. It is very easy to see how the result may be affected. Suppose some industrial disaster throws a great proportion out of work in August in one year, so as to increase the percentage of unemployed, we will say to 50, then when taking the average for ten years, that figure alone gives a rate of 5 per cent. in August, whereas the excess had nothing to do with the fact that August was the month concerned. If you take a sufficient number of years, however, those things will tend to equalize one another, and if we use the median instead of the arithmetic average extraordinary occurrences have little effect. For this reason it is best to estimate the period from the medians. In the end we shall not get a smooth curve for our averages, and may have to smooth that by a trigonometrical function, or by some other method.

Symptomatic Series.

We will now discuss the symptomatic curves; the top curve in Diagram XIII (male death-rate) will do as well as any as an illustration, for the method of dealing with this curve applies to a very great number of such curves. All statistics representing sociological phenomena that I have had experience of are symptomatic. Perhaps in very rare cases you will find no symptom, but in general there is a symptom; however remotely connected the figures are with the general progress of civilization, you will find there is some symptom up or down, or alternately up and down. In

general we may assume a symptom in all figures relating to human society. In dealing with such curves, we sometimes want to examine them in detail for a short period; but very often we are more concerned with the symptom, especially in forecasting events. In curve A in Diagram XIII there are considerable and rapid fluctuations, but there is also distinct optical evidence of a fall in the rate beginning between 1865 and 1870. The causes which produced the actual size of the ordinate are, of course, very many, and it is impossible to draw the line between those which tend to make a gradual permanent change, and those which tend to make a sudden temporary change. It is a question of degree and not of character, and for that reason alone it is impossible to give any theoretic solution for distinguishing the symptom from the small fluctuations, just as it is impossible to give any general solution to the interpolation problem. We have then to find an empirical solution, one that satisfies our immediate needs. It might appear best to draw a straight line, which on the whole shall differ from the observations as little as possible, and which could be determined by the method of least squares; this would assume a symptomatic tendency to equal increments or decrements in successive years. Or we might assume a parabolic curve or logarithmic curve. A recent American writer has assumed that a certain series can be represented by $y = kx^n$, the compound interest equation. But I think in general there is no reason to assume any definite algebraic law. The solution I should suggest—it is a commonplace one—is similar to that I have just suggested for the removal of the period. It is most easily understood by an example.

The figures in the following table are from the Registrar-General's Returns, or are calculated from the Statistical Abstract.

Years	Imports per Head.		Marriage Rate per 1,000		Death Rate of Males per 1,000		Death Rate of Females per 1,000		
	Amount	Deviation from Moving Average	Rate	Deviation from Moving Average	Amount	Deviation from Moving Average	Amount	Moving Average	Deviation from Moving Average
	£	x_1		x_2	£	x_3	£		x_4
1845	3·30	...	17·2	...	21·7	...	20·1	...	...
6	3·15	...	17·2	...	23·9	...	22·2	...	...
7	3·21	−·13	15·8	−·7	25·5	+1·4	23·9	22·6	+1·3
8	2·91	−·44	15·9	−·6	23·8	− ·3	22·2	22·6	− ·4
9	3·52	−·03	16·2	−·3	25·8	+1·9	24·4	22·4	+2·0
1850	3·97	+·19	17·2	+·4	21·4	−2·0	20·1	21·9	−1·8
1	4·14	−·16	17·2	·0	22·8	− ·6	21·2	21·8	− ·6
2	4·35	−·35	17·4	·0	23·2	+ ·1	21·5	21·5	·0
3	5·51	+·58	17·9	+·7	23·8	+ ·3	22·	21·8	+ ·4
4	5·51	+·17	17·2	+·1	24·4	+1·2	22·7	21·5	+1·2
5	5·16	−·64	16·2	−·7	23·5	+ ·4	21·7	21·4	+ ·3
6	6·16	+·30	16·7	+·2	21·3	−1·8	19·6	21·5	−1·9
7	6·66	+·65	16·5	·0	22·6	− ·3	21·1	21·2	− ·1
8	5·80	−·64	16·	−·7	23·9	+1·3	22·3	21·	+1·3
9	6·26	−·45	17·	+·4	23·3	+ ·4	21·5	21·2	+ ·3
1860	7·32	+·40	17·1	+·6	22·1	− ·8	20·3	21·	− ·7
1	7·50	+·05	16·3	−·4	22·7	− ·2	20·6	21·	− ·4
2	7·72	−·33	16·1	−·6	22·4	− ·8	20·5	21·2	− ·7
3	8·45	+·05	16·8	·0	24·1	+ ·4	21·9	21·5	+ ·4
4	9·26	+·40	17·2	+·2	24·9	+ ·8	22·5	21·8	+ ·7
5	9·06	−·06	17·5	+·4	24·5	+ ·3	22·	21·8	+ ·2
6	9·80	+·45	17·5	+·5	24·6	+ ·6	22·2	21·6	+ ·6
7	9·05	−·36	16·5	−·2	23·	− ·8	20·5	21·3	− ·8
8	9·60	+·06	16·1	−·3	23·1	− ·8	20·7	21·2	− ·5
9	9·54	−·30	15·9	−·4	23·6	·0	21·	21·	·0
1870	9·70	−·35	16·1	−·3	24·2	+ ·7	21·6	20·9	+ ·7
1	10·49	−·15	16·7	·0	23·9	+ ·6	21·3	20·7	+ ·6
2	11·13	+·28	17·4	+·4	22·6	− ·7	19·9	20·7	− ·8
3	11·54	+·35	17·6	+·5	22·4	− ·9	19·8	20·7	− ·9
4	11·39	+·04	17·	·0	23·6	+ ·6	20·9	20·3	+ ·6
5	11·39	−·08	16·7	·0	24·1	+1·3	21·4	20·1	+1·3
6	11·30	−·04	16·5	+·3	22·3	− ·6	19·6	20·2	− ·6
7	11·75	+·57	15·7	·0	21·7	− ·9	18·9	20·	−1·1
8	10·87	−·41	15·2	−·1	22·9	+ ·8	20·3	19·5	+ ·8
9	10·59	−·70	14·4	−·7	22·	+ ·3	19·6	19·2	+ ·4
1880	11·88	+·59	14·9	−·1	21·8	+ ·3	19·3	19·1	+ ·2
1	11·37	−·15	15·1	·0	20·	−1·1	17·8	18·7	− ·9
2	11·73	+·14	15·5	+·3	20·7	− ·1	18·5	18·5	·0
3	12·04	+·77	15·5	+·4	20·8	+ ·3	18·5	18·3	+ ·2
4	10·92	·0	15·1	+·1	20·9	+ ·2	18·5	18·4	+ ·1
5	10·30	−·26	14·5	−·2	20·3	− ·3	18·2	18·4	− ·2
6	9·63	−·62	14·2	−·3	20·6	+ ·4	18·5	18·1	+ ·4
7	9·90	−·47	14·4	−·1	20·2	+ ·3	18·1	17·8	+ ·3
8	10·51	−·04	14·4	−·3	19·2	− ·8	17·	17·8	− ·8
9	11·50	+·57	15·	·0	19·3	− ·9	17·2	17·9	− ·7
1890	11·22	+·05	15·5	+·3	20·8	+ ·6	18·3	17·9	+ ·4
1	11·52	+·34	15·6	+·4	21·5	+1·1	19·	18·1	+ ·9
2	11·12	+·14	15·4	+·1	20·	·0	18·	17·8	+ ·2
3	10·53	...	14·7	...	20·3	...	18·1	...	...
4	10·53	...	15·1	...	17·6	...	15·7	...	...
	Standard Deviations $\sigma_1 = ·386$		$\sigma_2 = ·37$		$\sigma_3 = ·830$		$\sigma_4 = ·803$		
	$\Sigma x_1 x_2 = 4·236$		$\Sigma x_1 x_3 = -3·207$		$\Sigma x_2 x_3 = -2·73$		$\Sigma x_3 x_4 = 30·38$		
	$r = \frac{4·236}{46\sigma_1\sigma_2} = ·65$		$r = \frac{-3·207}{46\sigma_1\sigma_3} = -·22$		$r = \frac{-2·73}{46\sigma_2\sigma_3} = -·19$		$r = \frac{30·38}{46\sigma_3\sigma_4} = ·99$		

Take the last group of figures, the death rate of females. I take the average of the first five death rates, 20·1, in 1845 to 24·4 in 1849, namely, 22·6, and place in the penultimate column at the middle of the period, namely, the year 1847. I begin again at the second year, 1846, and take the average for 1846-1850, namely, 22·6 again, and place that at 1848, the middle year of that period; and so on for 46 successive periods. Then on the Diagram XIII, I have represented that line of moving averages by the dotted line running through the continuous line. I think it is clear that that line offers one solution of the problem. In taking the average of any five years we are equally likely to include the ups and downs of their fluctations. If there was a regular period, if the fluctuations were five-yearly we should remove them entirely in five years, it would be the obvious time to take. If we were dealing with figures referring to industry and the period was ten years, ten years would be the most appropriate length of time to average, including as it would one contribution from each part of the fluctuation. If there is no regular period there is no rule to be given as to what number of years you shall take; it is a matter of convenience. If the five-yearly average gives you a curve with sharp angles and apparently random fluctuations, increase the number of years. It is most convenient to work with an odd number of years, for the middle of the period then coincides with the middle of one of the years; but, on the other hand, a period of ten years gives arithmetical facilities. This method may, I think, be left for consideration; I believe it will be seen that it offers a solution of the problem. To complete it, I recommend replacing the dotted line by a regular curve drawn very near to it, smoothing out any little fluctuations which are left. A curve thus drawn would fall from 1847 to 1858, and rise for about seven years and then fall, fairly rapidly, to about 1882, and more slowly afterwards. In the nature of things we cannot fix exact years for the end of the rise or fall. It is absolutely necessary to have some such method of measuring the symptom before you can base any argument as to the change in the quantity measured. That is very important. For example, the curve D, which represents imports, is a sharply fluctuating curve with a partial period. If, to take a particular date, we had in 1879 looked at the

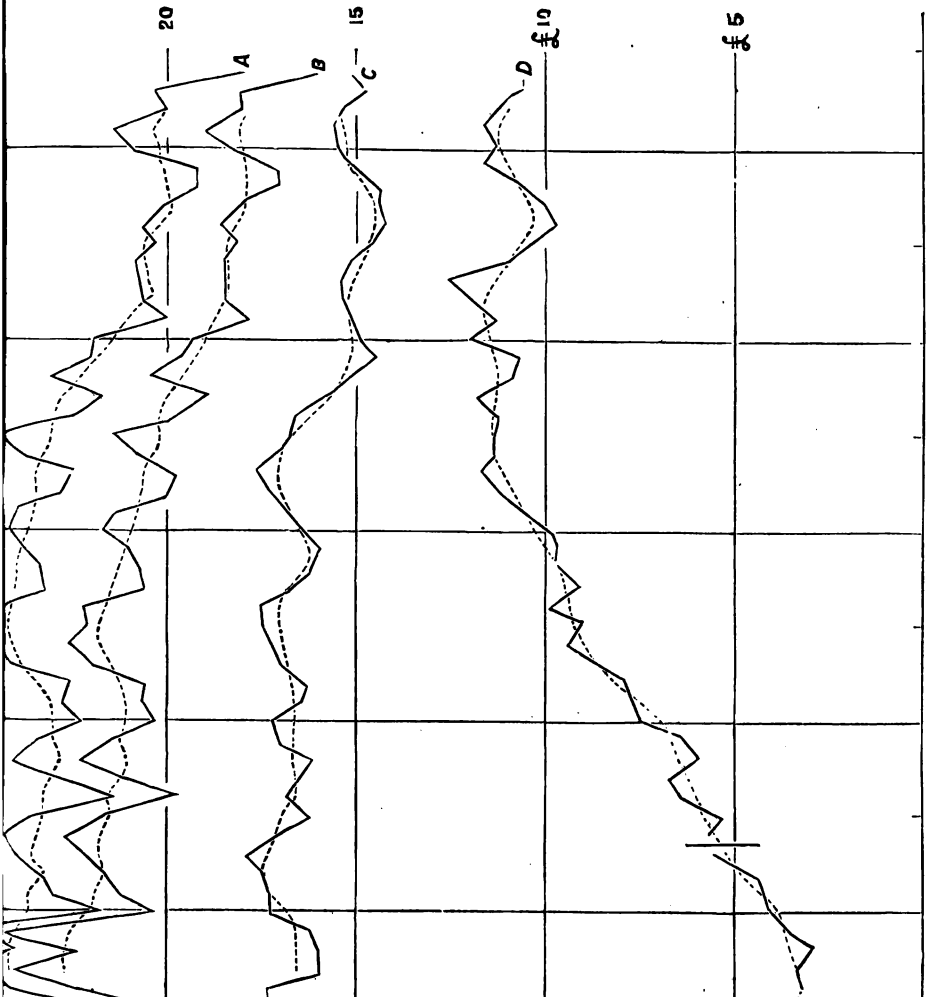


DIAGRAM XIII.

England.
- A Deaths per 1,000 living—Males.
- B „ „ „ „ —Females.
- C Persons married to 1,000 living.
- D Value of Imports per head of the population of the United Kingdom.

previous two years only we should have thought there was a rapid fall in the average imports; but if we looked at the history of the phenomenon we should have seen that it appeared to be only part of a minor fluctuation, and in 1882 we should have seen that average imports had been stationary on the whole for eight years.

It is not possible to say at the moment whether a fall is of a permanent nature or simply one of those little fluctuations which characterize the phenomenon throughout the half century. For instance, by 1903 we can perhaps judge of the tendency in 1900, but cannot judge of the current year because we have not enough information.

The deviations obtained by subtracting the instantaneous average from the figures for each year are given in the last column. The deviations for the first three groups of figures in the table are calculated on a similar method. These deviations should have some affinity to the curve of error. Great deviations should be rare compared with small deviations, and the occurrence of small and great deviations should have some such relation as the occurrence of great and small deviations in the curve of error; but the agreement is not likely to be close, for the deviations calculated here are not independent one of the other; they are bound together by the fact that the same number is used in forming five successive averages, while the curve of error assumes that the things are absolutely independent.

Correlation between Series.

That is a very rapid discussion of a rather wide subject, but I must lead on to the correlation between two sets of figures. If we were dealing with a curve with no symptom and no period, for instance, two sets of figures relating to the weather, $x_1, x_2 \ldots x_n$ representing the average temperature, $y_1, y_2 \ldots y_n$ representing the average wind velocity, the correlation between these two should be calculated as already described. If we were dealing with a periodic curve we should replace the periodic curve by its line of average before comparing it with another curve. If there is an irregular period, then I think we should proceed as if we had a symptomatic curve with no period. Of course, any two periodic curves with the same period are correlated. Any two sequences of events which are influenced by the annual

changes in the weather will give a strong degree of correlation quite independently of anything else. That is a quantity which in general will not be worth measuring, but when we come to very irregular periods such as those which we find in trade statistics, it is worth measuring the correlation even through the periods; because it is not so obvious, for instance, that all the fluctuations of exports are correlated with all the fluctuations of imports, and that the two together are correlated with the amount of employment.

A difficulty arises in dealing with many curves from the fact that the successive deviations year by year are not altogether independent. Many curves which deal with sociological phenomena have fluctuations each of which extends over several years, so that a rise in one year is more often followed by a rise in the next than by a fall. Other curves have the opposite character, that an excess in one year is followed by defects in the other; for instance, if there is a great death rate in one year we may expect a comparatively small one in the following; and this absence of independence should be kept in mind when we have to base arguments on the resulting correlation. But apart from this, we could treat the deviations from the moving line of averages as deviations whose correlation we can fairly calculate.

There is a very great difficulty in working out the correlation between symptomatic curves. If we do not take the deviation from the line of averages, but take the deviations from the average for the whole 50 years, any two symptomatic curves will show correlation. If we take two things which are absolutely disconnected, except that they are both phenomena arising in the progress of society, and work out the coefficient by the straightforward rule, we shall find there is some correlation. If two curves have short fluctuations which are correlated, but opposite symptoms, then owing to the symptom apart from the fluctuations there would be negative correlation, while owing to the fluctuations apart from the symptom there would be positive correlation; and when both are taken into account the correlation may be positive, zero, or negative. It is therefore necessary to treat the symptom separately from the short fluctuations. On the whole there is not much benefit in measuring the correlation coefficient for the symptoms; we should rather simply state that the symptom is say 15° upward in one case and 10° downward

in the other. The useful measurement of the correlation between two such curves is not that of the symptoms, but of the deviations.

Another question which arises very often in a practical way is, whether we should compare the deviation for the whole figures, say imports, with the deviation for the other, say the marriage rate, in the same year, or in the next year Can we correlate the imports of 1847 with the marriage rate of 1847, or should it be taken in comparison with that of 1848? That question will often occur, especially between marriage and birth rates. Mr. Hooker has suggested,* a suggestion which has been made independently in America, that we should work out the coefficients of correlation on the hypothesis of synchronism, and on alternate hypotheses that one event follows half or one year after the other, and see which correlation is the greatest. In this way we should get a series of correlation coefficients according to the dates we take.

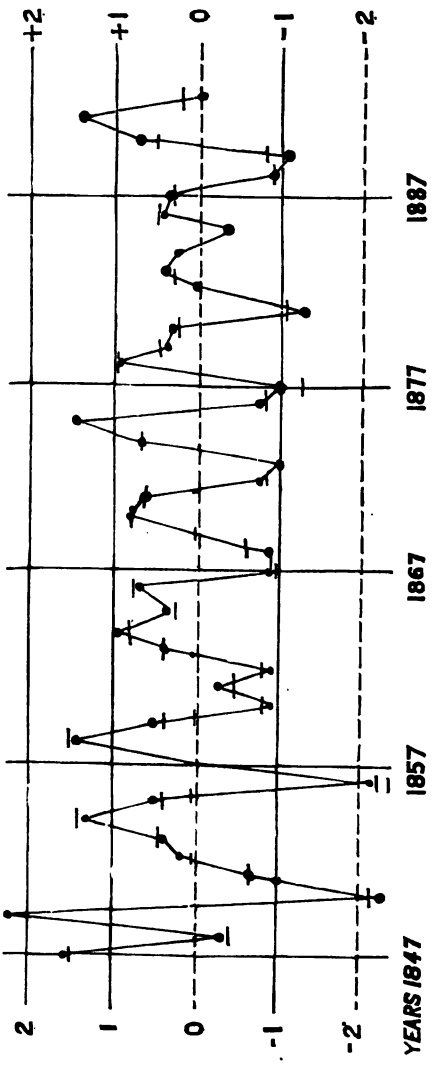
Before we proceed to measure correlation by mathematical formulæ we should observe it purely graphically; and the graphic representation of series will often suggest the existence of correlation, which can then be measured by the mathematical formula. The curves A and B in Diagram XIII are obviously closely correlated. In the curves B and C we cannot decide from the figure whether there is correlation or not; at any rate the evidence of correlation is not so great. In the curves B and D, I do not think we could decide from the figure as drawn, though we might perhaps from a figure drawn in a different way, whether there was correlation or not.

Let us proceed to discuss how to put two curves down so as to get optical evidence as to whether they are correlated or not. Instead of measuring figures as in Diagram XIII measure as in Diagram XIV. Plot out the deviations calculated on p. 79 above and below a base line representing zero; but before doing so it is necessary to choose the relative scales of the two quantities so as to have a definite relation the one to the other. There is no obvious way of comparing pounds sterling with one per thousand in the marriage rate. The way which naturally suggests itself

* *Journal of the Royal Statistical Society*, Sept. 1901. See especially pp. 490–1. I think that Mr. Hooker was also the first to publish a calculation of correlation based on deviations from a moving average.

DIAGRAM XIV.

Comparison of deviations of death rates from moving average.



and it is very useful for making the optical evidence of correlation vivid, is to represent the standard deviation for each group by unity on the vertical scale. The standard deviation for the death rate of males is ·830; of females it is ·803; so we represent the deviation ·830 for males and ·803 for females by the same vertical line. If we were doing the same thing for imports and male death rates, we

should represent by the same vertical scale ·386 of a pound sterling and ·83 death rate. This method has been applied in Diagram XIV, for the comparison of lines A and B in Diagram XIII. I have put in the death rate for males, represented by the zigzag line, but I found I could not enter the death rate for females in the same way and make the lines distinct, and therefore have drawn short horizontal lines to show the death rate for females year by year; the dots and lines representing the two series are in nearly every year close together. The optical evidence of correlation is very great indeed.

The illustration taken is of two series where the correlation is nearly perfect; in less perfect cases we can get evidence by noticing whether the maxima and minima occur at the same dates in the two series. For example, if we took the value of exports and percentage of unemployed we should find perhaps that the maximum of the one came at the same time as the minimum of the other throughout, and that would give strong optical evidence of negative correlation. A method of testing whether there was correlation or not, which would naturally suggest itself to anyone who has a small knowledge of probability, would be to see how often a positive deviation of the one agreed with a positive deviation of the other; how often like signs concurred and how often unlike signs concurred. If we wrote down 50 + and − signs at random and another 50 alongside, the chances of getting various numbers of agreements are easily calculated, and are in fact the successive terms in the expansion of $(\frac{1}{2}+\frac{1}{2})^n$; but that is not a good method, for it does not take into account one of the most important considerations, whether a great fluctuation of the one corresponds with a great fluctuation of the other or not. We should get equal evidence of correlation by this method when we had a resemblance of this sort in two curves where great fluctuations corresponded with great and small with small throughout, and when the correspondence was in sign only. Those things are obviously not of the same importance, and so the method of merely counting signs will not take us very far.

We will, then, proceed by the method of calculating correlation described above. Referring now to the table on p. 79, it should be remarked that the method of evaluating the value of imports changes at the year 1852; I had to

approximate before that year from the values of the exports, because the figures given by the Board of Trade before and after that date are calculated on different methods, and are not comparable. Otherwise the figures of total value of imports for home consumption to the United Kingdom are comparable. To my mind the imports are more significant than the exports; and also it seems to me absurd to add imports and exports; I do not think you can add them together any more than you can add bread to butter. I have taken the imports only, and, without criticising the figures in detail, I have divided by the gross population as given in the Statistical Abstract. Thus we get the amount per head given in the first column in the table. I have only intended to work to the second place of decimals. The death and marriage rates are taken from the Registrar-General's Report for 1895, which gives the figures for the previous 50 years. The standard deviations given at the bottom of the table are obtained by taking the square root of the sum of the squares of the 46 deviations given, divided by 46, as in the ordinary formula for standard deviation. The standard deviation is essentially an absolute quantity without sign.

I have calculated the coefficients of correlation between groups 1 and 2 (imports and marriage rate), between groups 1 and 3 (imports and death rates), between 2 and 3 (marriage and death rates), and between 3 and 4 (death rates for males and females). I have intended to choose cases where, *à priori* we might expect small correlation, no correlation, and great correlation. *À priori* we should expect correlation in the positive sense between imports and the marriage rate; not that increased imports cause an increase of the marriage rate, but the causes which produce prosperity are likely to have effect in increasing both imports and the marriage rate, the complexus of causes which decide the two things have something in common. The coefficient is ·65. The marriage rate and death rate have presumably very little in common. One certainly could not say to start with whether an increasing death rate would synchronize with an increasing or with a diminishing marriage rate. The correlation between the two is −·19. The correlation between the imports and the death rate is −·22. The correlation between the death rate for males and that for females is +·99; it is practically 1,

but the number 1 can only be obtained if there is an absolute proportion all through the scale, which there is not in this case.

Criterion of Significance of the Correlation Coefficient.

Now we are face to face with the question, What do these numerical values mean, and which of them are significant? It is clear that some such question arises, because if we write down two series absolutely at random and work out their formulæ the chances are very much against your obtaining zero, and there are heavy odds against obtaining a small number. Now the chance of obtaining a coefficient near zero increases with the number of terms. If we have two series, $u_1, u_2 \ldots u_n$ and $v_1, v_2 \ldots v_n$, measured from their averages, and we select a group of v's which are near to one another, the u's which will be their factors in forming the sum of the products are equally likely to be positive or negative; if we had an infinite number of these deviations their sum will be nothing; and the sum would tend to zero if we increased the number of terms, the actual deviation from zero being in inverse proportion to the square root of n, the number of terms. Hence the number of terms taken has much to do with the significance of the resulting coefficient of correlation, and we should expect that the quantity $\frac{1}{\sqrt{n}}$ would enter into the measurement of the significance of the coefficient of correlation. It is a little difficult to state and explain the measurement of the criterion of the significance; but it is absolutely necessary to make the attempt. Of the coefficients just given, the first and fourth are found to be significant, and the second and third not, when tested by the theoretical criterion.

Suppose we take two correlated groups, and that there is, as a matter of fact, a definite value for the coefficient of correlation; and then suppose we take 50 samples from each, that is to say, 50 pairs of events, we shall not naturally obtain exactly the coefficient of correlation that belongs to the whole groups. The chances are against obtaining exactly that result. Now, the deviations from the actual coefficient of correlation which are obtained by taking samples

and finding the correlation have a curve of frequency $y = \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}}$, where y is the probability that the coefficient obtained differs by x from the true coefficient, and c, the modulus $= (1-r^2) \cdot \sqrt{\frac{2}{n}}$, where r is the result obtained from the sample group, which consists of n pairs. The probable error in this curve of error is ·67 of $\frac{1-r^2}{\sqrt{n}}$.* For example, in the coefficient between imports and the marriage rate, $n = 46$, the calculated coefficient of correlation is ·65, and the probable error for its curve of frequency is $\cdot 67 \times \frac{1-(\cdot 65)^2}{\sqrt{46}} = \cdot 056$.

That is to say, from the calculation itself it is as likely as not that the actual coefficient is between ·65 + ·056 and ·65 − ·056. The chance of the true coefficient being as much as the modulus, namely ·115, distant from the calculated ·65 is shown by the table of the error function to be only 16 in 100; the chance of it being so far from ·65 as to be actually zero are infinitesimal, for in the curve of error the cases where the deviation is as much as six times the modulus are practically non-existent. So that we have overwhelming evidence, if our general principle of calculation is correct, of correlation between the first and the second columns, and the most probable value of that correlation is two-thirds. In other words, the standard deviation of imports being £·387, and the standard deviation of the marriage rate ·37, the most probable deviation of the marriage rate is $+\frac{2}{3}$ of ·37 = ·24, when we find a deviation in imports of +£·387, and so on in proportion. This statement should be connected with the graphic measurement of correlation discussed on p. 70. In the second case of correlation, that between imports and the male death rate, where the coefficient is −·22 the probable error by the method just described is ·09. That is to say, our calculation means that it is as likely as not, from our evidence, that the correlation between these two series is between −·13 and −·31; the chance that the real correlation is zero or positive is quite perceptible. The chance from the table of the error

* *See* Pearson, in *Royal Soc. Trans.*, A. 175, p. 265; and correction, in *Royal Soc. Proceedings*, Oct. 18th, 1897; also Yale, in *Statistical Journal* 1897, p. 847.